Modelling Covalent Modification of Cysteine Residues in Proteins

by

© **Ernest Awoonor-Williams**

A thesis submitted to the School of Graduate Studies in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Department of Chemistry

Memorial University

May 2020

St. John's, Newfoundland and Labrador, Canada

Abstract

Cysteine is a unique amino acid because of the chemical reactivity of its thiol (–SH) side chain. For that reason, cysteine serves several essential roles in biochemistry, and its reactivity is critical for the catalytic activity of several biological enzymes. This significance of cysteine residues has been exploited in designing covalent-modifier drugs, particularly kinase inhibitors, which have proven to be successful cancer chemotherapeutic agents in recent years. The reactivity of cysteine thiol group is complex, but a measure of its acidity or pK_a is a strong determinant of its reactivity towards druggable targets—and can help guide the selection of appropriate druggable targets for covalent modification. Relatively few experimental pK_a's of cysteine residues in proteins have been reported, and methods for the computation of cysteine pK_a's have received little attention.

This thesis presents studies undertaken to investigate the reactivity and covalent modification of cysteine residues in proteins. The introductory chapter lays the groundwork that becomes the basis for subsequent chapters in the thesis. This chapter provides a general introduction to covalent modification and the techniques used to investigate the biophysical properties of residue-specific nucleophilic targets for covalent modification. The first two chapters following the introduction are focused on predictive pK_a assessments and validation studies on different computational methods in accurately calculating experimental cysteine pK_a's. In the latter chapters, advanced computational and multiscale methods are adopted to investigate the reactivity of druggable cysteines in protein kinases commonly implicated in diverse clinical indications, as well as model all the steps in the covalent modification mechanism of a kinase target. The thesis concludes by providing a concise summary of the research findings and future directions stemming from the work.

The fundamental studies presented herein expand our current knowledge of modelling the covalent inhibition of druggable cysteines in enzyme targets and could go a long way to inform drug design and discovery.

Dedicated to my family, especially my parents.

Acknowledgement

First and foremost, I thank my supervisor and mentor, Prof. Chris Rowley, for teaching and inspiring me to be a better scientist and researcher. I admire the wealth, breadth, and depth of your knowledge. Thank you for your guidance and mentorship during my graduate degree. In addition, thank you for your prompt feedback and multiple rounds of assessment of all my research work, especially this document.

I also thank my supervisory committee members: Prof. Travis Fridgen & Prof. Chris Kozak, for their advice, guidance, and assessment of my research work.

I would also like to acknowledge my thesis examiners: Prof. James Gauld, Prof. Lindsay Cahill, and Prof. Stefan Wallin, who provided helpful suggestions to improve the final version of this thesis.

To all the past and present members of the Rowley Group during the past six years, I am grateful for the many fond memories shared and insightful discussions. You have made this experience a pleasant and enjoyable one. Special mention to Jessica Besaw (MSc, 2015), Kari Gaalswyk (MSc, 2016), Maxime Lecocq (French Exchange Student, 2016), Mohamad Mohebifar (MSc. 2018), Fatima Sajadi (MSc, 2018), Archita Adluri (B.Sc. Hons, 2018), and Evan Walters (B.Sc. Hons, 2018). I would also like to thank Shae-Lynn Lahey, Sheyla Montero, Nazanin Rezajooei, and Tù Nguyên Thiên Phúc for being kind and receptive to reading various sections of this document.

Special thanks to Abd Al-Aziz A. Abu-saleh of the Poirier Group who has been a very kind and thoughtful friend, as well as an awesome workout buddy. I have enjoyed our thorough discussions about science, research, and life in general. Your in-depth level of knowledge about different types of Gaussian calculations is greatly admired.

I would also like to extend my gratitude to my friends, colleagues, Department of Chemistry support staff and faculty.

I would also like to acknowledge my parents and siblings for their constant love and support over the years. In particular, my parents for sacrificing a lot towards my education and always believing in me. I will forever remain grateful!

Lastly, I thank the Natural Sciences and Engineering Research Council (NSERC) of Canada for funding through the Vanier Canada Grauate Scholarship Program. I also thank Atlantic Computational Excellence network (ACEnet) for an Advanced Research Computing Fellowship and the School of Graduate Studies of Memorial University of Newfoundland for a graduate fellowship.

Computational resources were provided by Compute Canada through the Calcul Québec, SHARCNET, and ACEnet consortia, and through the Center for Health Informatics and Analytics (CHIA) of the Faculty of Medicine at Memorial University of Newfoundland.

Statement of contribution

A significant portion of the work reported in this thesis includes text, figures, and tables from four published manuscripts. I am the first author in all these manuscripts.

Chapter 1:

The first chapter provides a general overview of modelling covalent-modifier drugs and the computational techniques used to access the reactivity of active-site nucleophilic side chain groups in proteins that bind to these drugs. The bulk of this chapter was prepared by me with small sections adapted with permission from:

Awoonor-Williams, E.; Walsh, A. G.; Rowley, C. N. Modeling Covalent-Modifier Drugs. *Biochim. Biophys. Acta, Proteins Proteomics*, **2017**, 1865, 1664–1675.

I was responsible for writing sections of this manuscript. I performed the molecular docking simulations and generated publication-quality figures. I received guidance and assistance from Dr. Rowley, who also contributed to writing and generating figures for sections of the manuscript.

Chapter 2:

This chapter evaluates the performance of computational methods in accurately predicting the experimental pK_a's of cysteine residues in a test set of proteins. The chapter is based on the following publication in The Journal of Chemical Theory and Computation:

Awoonor-Williams, E. and Rowley, C. N. Evaluation of Methods for the Calculation of the pK_a of Cysteine Residues in Proteins *J. Chem. Theory Comput.*, **2016**, 12 (9), 4662–4673.

I performed all the simulations and generated the figures and tables for this manuscript. I also wrote the first draft of the manuscript and received assistance from Dr. Rowley prior to submission of the manuscript.

Chapter 3:

This chapter employs free energy perturbation and multiscale simulation methods to assess the parameters of cysteine thiolate used by the popular CHARMM and Amber force fields in biomolecular simulations. The chapter is based on the following publication in The Journal of Chemical Physics:

Awoonor-Williams, E. and Rowley, C. N. The Hydration Structure of Methylthiolate from QM/MM Molecular Dynamics *J. Chem. Phys.*, **2018**, 149, 045103–8.

I wrote sections of the manuscript and performed both the free energy perturbation and quantum mechanical simulations. I also generated figures used in the manuscript. Dr. Rowley assisted in writing sections of the manuscript and performing simulations for the manuscript.

Chapter 4:

This chapter employs rigorous computational approaches to calculate the acidity of druggable cysteines in protein kinases that are commonly implicated in diverse clinical indications. Much of this chapter is based on a publication in The Journal of Chemical Information and Modeling:

Awoonor-Williams, E. and Rowley, C. N. How Reactive are Druggable Cysteines in Protein Kinases? *J. Chem. Inf. Model*, **2018**, 58(9), 1935–1946.

The entirety of the work reported in this manuscript was carried out by me. I performed all the simulations and was the primary contributor to the manuscript. Dr. Rowley was a secondary contributor to the manuscript.

Chapter 5:

In this chapter, a multiscale computational approach is undertaken in an effort to model all the steps in the covalent binding process of cysteine-targeting covalent inhibitors. The entirety of the work reported in this chapter was prepared by me.

Chapter 6:

This chapter provides a brief summary of the results presented with possibilities for future work in the field. The entirety of the work reported in this chapter was prepared by me.

Table of contents

List of tables

List of figures

List of abbreviations

AIMD	*Ab initio* Molecular Dynamics
ATP	Adenosine Triphosphate
B3LYP	Becke, 3-parameter, Lee-Yang-Parr
BTK	Bruton's Tyrosine Kinase
CCSD(T)	Coupled Cluster Single-Double and perturbative Triple
CpHMD	Constant-pH Molecular Dynamics
Cys	Cysteine
DFG	Aspartate-Phenylalanine-Glycine
DFT	Density Functional Theory
EGFR	Epidermal Growth Factor Receptor
FEP	Free Energy Perturbation
FIRES	Flexible Inner Region Ensemble Separator
Fs-IFP	Functional-site Interaction Fingerprint
MCCE	Multi-Conformation Continuum Electrostatics
MD	Molecular Dynamics
neMD/MC	nonequilibrium Molecular Dynamics/Monte Carlo
NMR	Nuclear Magnetic Resonance
NpT	Isothermal-isobaric Ensemble
NVT	Isothermal-isochoric Ensemble
ONIOM	Our own N-layered Integrated molecular Orbital and molecular Mechanics
PBE	Perdew-Burke-Ernzerhof
PCM	Polarized Continuum Model
PDB	Protein Data Bank
PME	Particle Mesh Ewald
QM/MM	Quantum Mechanics/Molecular Mechanics
REMD	Replica-Exchange Molecular Dynamics
RESP	Restrained Electrostatic Potential
RETI	Replica-Exchange Thermodynamic Integration
RDF	Radial Distribution Function
RMSD	Root-Mean-Square Deviation
SAPT	Symmetry-Adaptive Perturbation Theory
SASA	Solvent Accessible Surface Area
TCI	Targeted Covalent Inhibitor
TI	Thermodynamic Integration
vdW	van der Waals force
WHAM	Weighted Histogram Analysis Method

List of symbols

Å	angstrom, 10^{-10} m
°C	degree Celcius
fs	femtosecond, 10^{-15} s
ΔG	relative Gibbs energy
$\Delta G^{\ddagger}$	relative Gibbs energy of activation
$\Delta\Delta G$	relative Gibbs energy difference
ϵ	Lennard-Jones well depth
σ	Lennard-Jones radii
h	Planck's constant, $6.626 \times 10^{-34} J \cdot s$
$\mathcal{H}$	Hamiltonian
IC_{50}	half maximal inhibitory concentration
k	rate constant
$k_{inact.}$	inactivation rate constant
K	kelvin
k_B	Boltzmann constant, $1.381 \times 10^{-23} J \cdot K^{-1}$
K_i	equilibrium constant
$kcal/mol$	kilocalorie per mole
kJ/mol	kilojuole per mole
ln	natural log
ns	nanosecond, $10^{-9} s$
p	momenta
Pa	pascal, $kg \cdot m^{-1} \cdot s^{-2}$
pK_a	equilibrium acidity
pK_{intr}	intrinsic pK_a
ps	picosecond, $10^{-12} s$
q	partial charge
R	molar gas constant, $8.314 J \cdot K^{-1} \cdot mol^{-1}$
T	temperature
$\mathcal{V}$	potential energy function
λ	reaction coordinate

"A theory that you can't explain to a bartender is probably no damn good."

— Sir Ernest Rutherford

1 Introduction

Sections of this introductory chapter has been published as a review: Awoonor-Williams, E.; Walsh, A. G.; Rowley, C. N. Modeling Covalent-Modifier Drugs. *Biochim. Biophys. Acta, Proteins Proteomics*, **2017**, 1865, 1664–1675.

Contents

1.1 Introduction

The general mechanism for the inhibition of an enzyme or receptor by a small molecule drug is for the drug to bind to the protein, attenuating its activity. The canonical mode by which a drug will bind to its target is through non-covalent interactions, such as hydrogen bonding, dipole–dipole interactions, and London dispersion interactions. Kollman and coworkers estimated that small molecules that bind to proteins through non-covalent interactions have a maximum binding affinity of 6.3 kJ/mol per non-hydrogen atom,[1] so these binding energies are generally sufficiently weak for the binding to be reversible. This establishes a measurable equilibrium between the bound and unbound states.

A sizable class of drugs bind to their targets by an additional mode; a covalent bond is formed between the ligand and its target. These drugs contain a moiety that can undergo a chemical reaction with an amino acid side chain of the target protein, covalently modifying the protein. Dissociating this covalent modifier from the target requires these protein–ligand bonds to be broken. In the cases where the dissociation is strongly exergonic, the equilibrium will lie so far towards the bound state that the inhibition is effectively irreversible. Some covalent protein–ligand reactions are only weakly exergonic, so covalent modification can be reversible in some instances.[2] These ligands also interact with the protein through conventional non-covalent intermolecular forces, so the total binding energy in the covalently-bound state results from both covalent and non-covalent interactions. Refs. [3–10] are recent reviews on covalent-modifiers in drug discovery.

The covalent modification of proteins can involve many types of chemical motifs in the inhibitor and involve a variety of amino acids. Catalytic residues in the active site often have depressed pK_a's, so they are more likely to occupy the deprotonated state that is reactive towards electrophiles. Serine residues that serve as a Brønsted acid in an enzymatic catalytic cycle have been a popular target. Two of the most famous covalent modifier drugs, aspirin and penicillin (Figure 1.1), act by acylating this type of active site serine residue. Aspirin targets Ser530 of cyclooxygenase and penicillin targets the active site serine in DD-transpeptidase enzymes to inhibit the growth and survival of bacterial cell wall. Catalytic serine, cysteine, threonine, and lysine residues have all been targeted by covalent modifiers.[11,12]

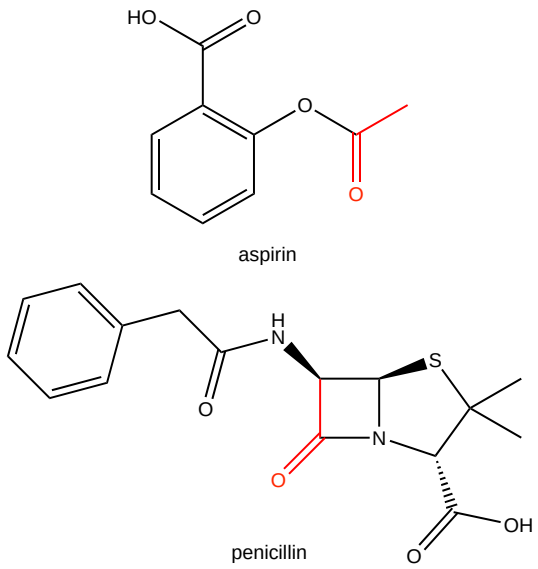

Figure 1.1: Aspirin and penicillin are early examples of covalent modifier drugs. These drugs bind to their target by acetylating a catalytic serine residue. The reactive moiety is drawn in red.

In recent years, there have been extensive efforts to develop covalent modifier drugs that undergo reactions with the thiol group of non-catalytic cysteine residues. Cysteines are relatively rare amino acids, comprising only 2.3% of the residues in the human proteome.[13] This limits the number of off-target reactions that are possible. These non-catalytic residues are less likely to be conserved within a family of proteins, which creates opportunities to select for a specific target in a large family of proteins. Although this type of covalent modification does not inactivate the catalytic residues directly, the covalent linker serves to anchor the drug binding site and achieve a stronger binding affinity. Large-scale screens have shown that reactive fragments have unexpectedly high specificity for individual proteins, suggesting that covalent modifiers have a lower risk of promiscuity than had been previously assumed.[14–16]

These advantages must be balanced against the drawbacks associated with covalent protein–ligand binding. Covalent protein–ligand adducts are believed to trigger immune responses in some cases.[17] Further, the inhibitor must be carefully tuned so that it will only bind irreversibly to its target because irreversible inhibition of an off-target receptor could result in adverse drug reactions. The chemical reactivity of the electrophilic functional group of the ligand (a.k.a., "warhead") also creates the potential that the inhibitor will be chemically degraded in an inactive form through

metabolism or other types of chemical reactions before it reaches the target. This constrains the reactivity of the warhead.

To develop drugs of practical use that have the advantages of covalent modification but avoid the disadvantages, researchers have developed targeted covalent inhibitors (TCIs). Typically, these compounds have a non-covalently binding framework that is highly selective for the target. For example, covalent modifier ibrutinib (Figure 1.2) shares the aminopyrimidine scaffold that has been successfully employed in the development of Bruton's tyrosine kinase-selective non-covalent inhibitors. The reactive warhead is an acrylamide, which is a moderately-reactive electrophile. Thiol-Michael additions are generally only weakly exergonic, so these additions are often reversible.[18] This allows the inhibitor to dissociate if it reacts with a thiol of a protein other than its target. High selectivity is achieved by the combination of selective non-covalent interactions and the additional strength of the covalent interaction between the warhead and a complementary reactive amino acid.

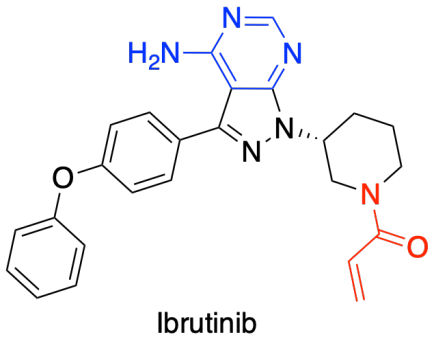

Figure 1.2: Ibrutinib is an example of a targeted covalent inhibitor used for the treatment of B cell cancers. The aminopyrimidine scaffold is highlighted in blue and the reactive acrylamide warhead is indicated in red.

1.1.1 Physical Parameters of Covalent Modification

The strength of the binding of a non-covalent inhibitor to its target can be quantified by the equilibrium constant, K_I, for the association of the ligand (inhibitor, I) and its target (enzyme, E) to form a protein–ligand complex (E·I). This association can also be defined in terms of the Gibbs energy of binding through the relation $\Delta G_{non-covalent} = -RT \ln K_I C^\circ$, where C° is the standard state concentration.[19] The binding of a covalent modifier involves additional steps. The protein–ligand complex

(E·I) undergoes reaction to form the covalent adduct (I–E), Figure 1.3. The rate of this process is characterized by its rate constant, $k_{inact.}$. In some cases, this reaction is reversible and the covalent adduct can revert to the non-covalent protein–ligand complex with the rate constant $k_{-inact.}$. If the reaction is strongly exergonic, $k_{inact.}$ will be much larger than $k_{-inact.}$, so the binding will effectively be irreversible. The total binding energy of the ligand results from both the covalent and non-covalent protein–ligand interactions (i.e., $\Delta G_{covalent}$ and $\Delta G_{non-covalent}$).

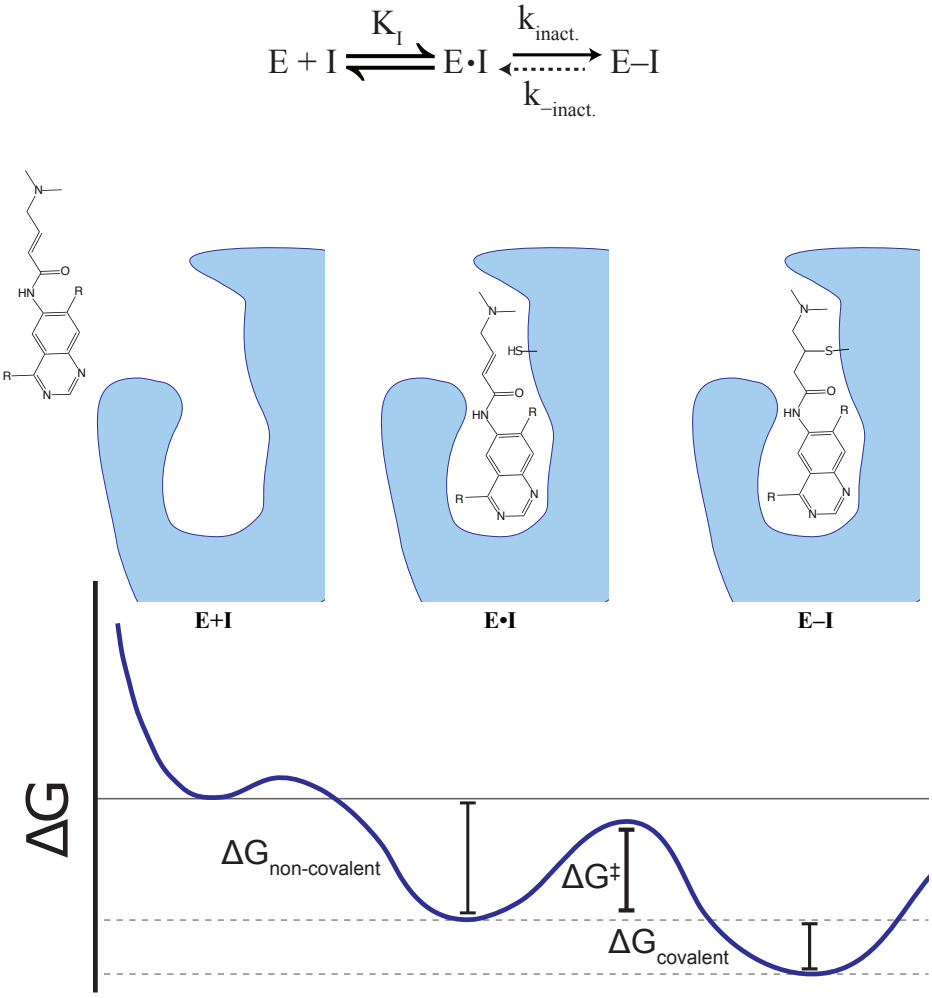

Figure 1.3: Schematic binding profile for the formation of a covalent protein–ligand complex. In this example, afatinib binds non-covalently to the active site of epidermal growth factor receptor (EGFR) kinase (E·I), then undergoes a chemical reaction with the thiol group of Cys-797 to form a covalent thioether adduct (E–I).

The rate at which an inhibitor reacts with the target ($k_{inact.}$) can be calculated using transition state theory. Conventional transition state theory is the simplest and

most widely used theory, which relates the rate of reaction to the Gibbs energy profile along the reaction coordinate[1]. Using transition state theory, the rate of reaction can be calculated by the Gibbs energy of activation ($\Delta G^{\ddagger}$ of the rate limiting step[21]),

$$k_{TST} = \frac{k_B T}{h} \exp\left(\frac{-\Delta G^{\ddagger}}{k_B T}\right). \tag{1.1}$$

The mechanism of covalent modification can be complex and involve multiple reaction steps. For example, the mechanism of covalent modification of a cysteine by a Michael acceptor involves the deprotonation of the thiol to form a thiolate, formation of an enolate intermediate, and protonation of the enolate to form the thioether product (Figure 1.4). A comprehensive model for covalent modification requires the calculation of $\Delta G_{non-covalent}$, $\Delta G_{covalent}$, as well as the rate-limiting barriers of the chemical reaction ($\Delta G^{\ddagger}$). The rates of binding, unbinding, inactivation, and activation govern the drug residence time, which has been proposed to be a better determinant of *in vivo* pharmacological activity than the binding affinity.[22–24]

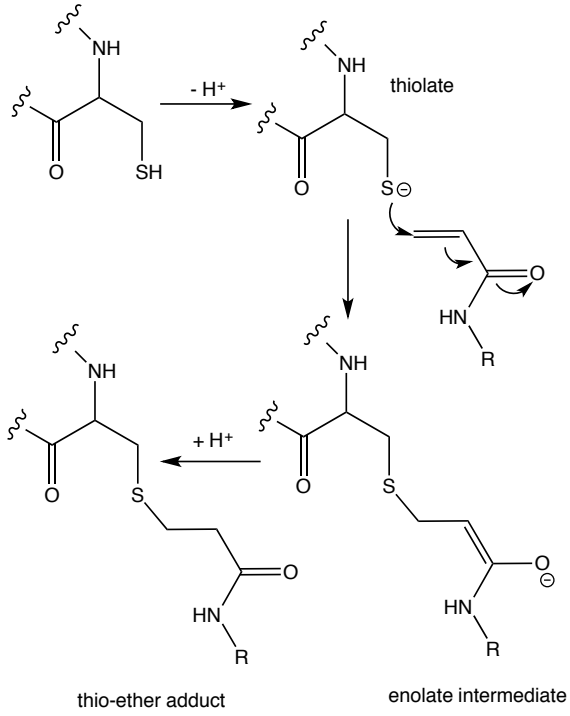

Figure 1.4: The mechanism of the addition of an acrylamide warhead to a cysteine thiol.

[1] A discussion of the limitations of this model in enzymatic reactions is available in Ref. 20.

1.2 Determination of the pK_a of Targeted Residues

Generally, the first step in the mechanism for covalent modification of cysteines, lysines, and serines is their deprotonation to yield their more reactive form (Figure 1.5). In the case of the modification of a cysteine residue by an electrophile, the thiol group of the amino acid side-chain must be deprotonated to form the reactive thiolate nucleophile. The stability of the thiolate is thus a significant parameter for the inhibition of a target site by a drug molecule. The equilibrium between the thiol and thiolate states of a cysteine residue in a protein is defined by its pK_a. Cysteines with low pK_a's are more likely to exist in their reactive thiolate state, so they will be more susceptible to covalent modification by electrophilic inhibitors.

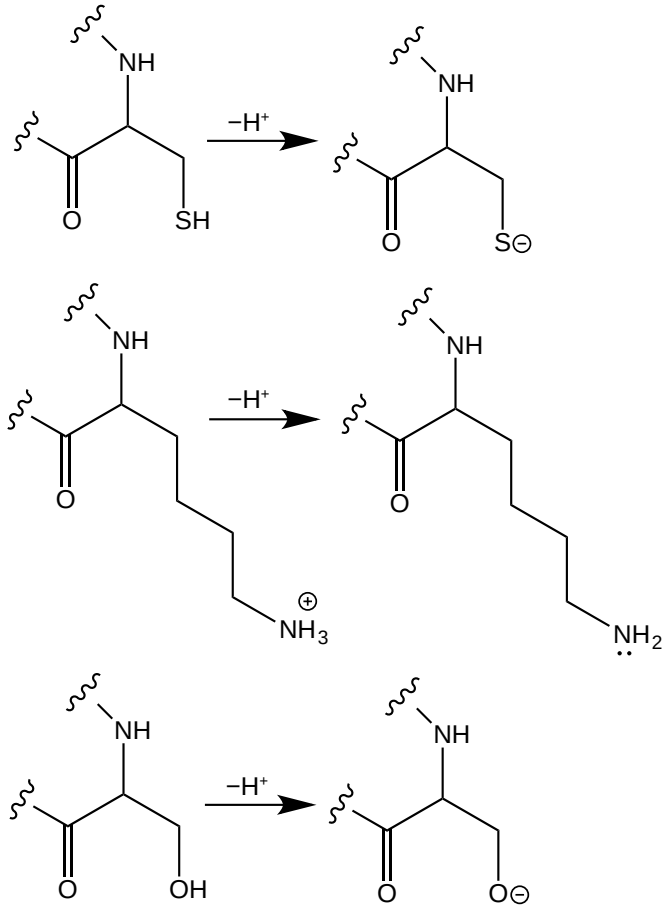

Figure 1.5: Deprotonation reactions involving cysteine, lysine, and serine. Covalent modification of these residues typically involves reaction in their deprotonated, nucleophilic states.

1.2.1 Factors Affecting the pK_a of Ionizable Residues

Typically, the pK_a's of ionizable residues in a protein will vary due to intrinsic differences in bonding with functional groups of the side chain (e.g., $-C(=O)OH$ vs. $-NH_3^+$, etc.) and the nature of the environment where the ionizable residue resides. As a general rule, the pK_a of ionizable residues in a positively charged environment will be lowered from their intrinsic pK_a values. For example, placing an Asp or Glu residue near positively charged Lys residues stabilizes the negative form of the Asp and Glu residues, lowering their pK_a.[25] Conversely, in a negative electrostatic potential environment, the pK_a will be increased. The intrinsic pK_a (i.e., pK_{intr}) of an ionizable residue is the pK_a of that residue when all other titratable groups in the protein are fixed in their neutral or reference states, Table 1.1. For ionizable amino acid residues, the intrinsic pK_a is determined by synthesizing a pentapeptide with blocked neutral terminal groups, where the ionizable residue of interest is positioned between two alanine side chains. This provides the pK_a of the solvent-exposed amino acid residue without electrostatic interaction effects from other neighboring ionizable residues with charged functional groups. The pK_a of an amino acid in a protein can deviate from this intrinsic value due to differences in its intermolecular interactions in its folded state.

Table 1.1: Intrinsic pK values of ionizable groups in proteins

Ionizable Group	**pK_{intr}**
α-amino	8.0
Asp	3.9
Glu	4.3
His	6.5
Cys	8.6
Tyr	9.8
Lys	10.4
Arg	12.3
α-carboxyl	3.7

N.B., pK_{intr} values reported above were determined in alanine pentapeptides with blocked termini [Ac-$(Ala)_2$-Cys-$(Ala)_2$-NH_2], and were taken from Refs. [26] and [27].

The standard pK_a of a free cysteine in solution is 8.6,[27] but pK_a's of cysteines have been reported to range from 2.9 to 11.1.[28–30] This broad range results from the intermolecular interactions that the thiol and thiolate states of the cysteine experience

inside the protein. Catalytic cysteines in enzymes like cysteine proteases tend to have nearby cationic residues, like histidine or asparagine, which lowers their pK_a's by stabilizing the thiolate state of the cysteine.[28] Conversely, the thiolate state of the cysteine residue will experience repulsive interactions with nearby anionic residues, raising the pK_a. Amino acids buried in hydrophobic pockets of the protein can also have elevated pK_a's because they do not experience stabilizing interactions with water molecules.

Three competing interactions are generally responsible for the environmental perturbation in pK_a of ionizable residues, namely: hydrogen bonding, desolvation effects, and electrostatic interactions. Hydrogen bonding tends to lower the pK_a of ionizable residues in a deprotonated state and raises the pK_a of ionizable residues in a protonated state. In proteins, hydrogen bonding generally confers 1–2 kcal/mol to structure stability.[31] Regarding desolvation effects, the neutral state of ionizable residues will be favoured when buried in the hydrophobic core of the protein. This is primarily due to lack of water in the protein interior, which has the ability to form favourable interactions with ionizable residues—stabilizing their charged states. As a result, the pK_a's of cysteine residues will be elevated when buried in the protein interior. Also, electrostatic interactions between charged groups when attractive and favourable, can help stabilize ionizable residues in proteins. This has the effect of lowering the pK_a's of acidic residues while raising the pK_a's of basic residues.

1.2.2 Methods for pK_a Determination

Experimental Methods

A number of experimental methods exist for determining the pK_a's of molecules, although the choice of technique is dependent on the sample and matrix under investigation.[32] For ionizable residues in proteins, potentiometric titration, quantitative mass spectrometry, and nuclear magnetic resonance (NMR) are among the most common experimental methods. A vast majority of these methods require the measurement of a physical parameter as a function of pH. For instance, in NMR titrations the mean chemical shift of an assigned proton near the ionization site is measured at varied pH levels. The pK_a value can then be determined by locating a point of inflection on the plotted titration curve. This approach of pK_a determination is described by

the Henderson–Hasselbalch equation, (Eqn. 1.2). The pK_a's of titratable sites will have to be independent and differ significantly from one another (i.e., > 2 pK units) in order to be accurately determined by this approach. The pK_a value is generically equal to the pH at which the protonation probability $\langle x \rangle$, given by the expression in Eqn. (1.3), is 0.5. For proteins and biomolecules where individual titratable residues interact and couple with each other, the titration curves can be considerably more complex and highly irregular. This effect of cooperativity in the interactions between titratable sites in proteins can result in ill-defined inflection points on titration curves—making it difficult to decipher pK_a values. Also, the variations in protein pK_a's due to environmental and conformational effects can further perturb the pK_a values.

$$pK_a = pH + \log\left(\frac{[\text{Conjugate Acid}]}{[\text{Conjugate Base}]}\right) \tag{1.2}$$

$$\langle x \rangle = \frac{10^{pK_a - pH}}{1 + 10^{pK_a - pH}} \tag{1.3}$$

Computational Methods

The complications in experimental protein pK_a determination have led to the emergence of a number of computational techniques for protein pK_a prediction.[33–41] Practically, these methods are simpler than experimental pK_a techniques because the rather cumbersome and time-consuming process of protein expression and synthesis is avoided. These methods provide pK_a estimates of titratable sites of interest using the three-dimensional protein structure, traditionally determined via X-ray crystallography or NMR. The pK_a estimates are based on the relative stability of titratable states and are obtained by computing the differences in free energy between charged and neutral forms of the amino acid of interest. The generic computational strategy involves the estimation of a pK_a shift (i.e., ΔpK_a), which depends on the environment of the residue. The pK_a shift is combined with a reference or intrinsic pK_a value to obtain the pK_a of the residue of interest. This can be formally expressed as:[42]

$$pK_a(\text{residue}) = pK_a(\text{reference}) + \Delta pK_a(\textit{solvent} \rightarrow \textit{protein}) \tag{1.4}$$

where pK_a(reference) refers to the reference pK_a value of the residue in solution (i.e., pK_{intr}). ΔpK_a is the pK_a difference of removing the residue from pure solvent and embedding it into the protein; a resultant term arising from the interactions experienced by the ionizable residue in its new environment.

The majority of computational pK_a prediction methods are based on an electrostatic approach, with calculations designed to accurately describe and estimate both the Coulombic interaction and solvation energy terms of ionizable residues within the model structure. These models are based on either a macroscopic or microscopic structural framework. Under the macroscopic framework, the system is modelled as a continuum dielectric with relevant interactions accounted for by solutions to macromolecular electrostatics equations, like the Poisson–Boltzmann Equation (PBE). Microscopic electrostatic methods on the other hand, are based on an atomistic model framework and attempt to treat all interactions in an *ab initio* quantum mechanical (QM) fashion. Macroscopic continuum electrostatic models were among the earliest methods used,[43] although the approximations incorporated in their methodology limit their accuracy. Since this time, these models have been continually improved, and some methods that make use of an explicit solvent representation perform well for predicting the pK_a's of aspartic and glutamic acid residues.[42,44] More recently, methods established on the basis of experimental parameterization of large datasets of protein pK_a's have also been developed.[38,41] These so-called empirical methods, are simply mathematical models trained on optimized parameters from large databases of experimental pK_a values. An example is the widely used program, PROPKA, developed by Jensen and coworkers.[38,45] PROPKA can predict side-chain pK_a's of Asp and Glu residues with a root-mean-square deviation of 0.79.[45] The method estimates pK_a shifts based on a physical description of desolvation effects, hydrogen bonding, and Coulombic interactions between charged groups.

1.2.3 Challenges in pK_a Calculation of Targeted Residues

To accurately calculate the pK_a's of targetable ionizable residues in a protein, one needs to consider the solvation environment, the protonation state of nearby titratable residues, and pH-induced conformational or structural changes. In addition, pH-dependent interactions that lead to favourable hydrogen bonding, inter-residue electrostatic interactions, and salt-bridges in proteins need to be accounted for to

yield accurate pK_a's. Methods that are capable of describing the coupling between conformational changes and variation in protonation/deprotonation events can significantly improve the accuracy of predictive pK_a calculation methods.

Despite the plethora of predictive pK_a methods, variations exist in the level of accuracy, performance, and cost of these algorithms in calculating ionizable residue pK_a's. Accurate calculation of the pK_a's of amino acid side chain groups in proteins has been a long-standing challenge in computational biophysics. Benchmark studies on the accuracy and performance of standard predictive pK_a methods suggest the need for a more thorough assessment and comparison of methods—an approach which could significantly enhance their accuracy and predictive capabilities. A significant portion of the work reported in this thesis is focussed towards addressing some of the challenges of existing computational methods in accurately predicting targetable residue pK_a, particularly cysteine.

The methods for prediction of the pK_a of cysteine residues in proteins are less established and have received very little attention. This is explored in-depth in Chapter 2, where a benchmark assessment of different computational methods is performed for predicting experimental cysteine pK_a's in a test set of proteins. Computational methods that employ either an implicit or explicit representation of solvent water molecules in the pK_a calculation were evaluated to determine their predictive accuracy. In Chapter 3, the cysteine thiolate parameters in the popular Amber and CHARMM molecular mechanical protein force fields were evaluated using advanced multiscale methods in an effort to validate and ascertain their use for biomolecular simulations, including pK_a calculations. The hydration structure of a model thiolate (i.e., methylthiolate) was calculated using *ab initio* molecular dynamics methods and compared with the structures predicted by molecular mechanics force field models. In Chapter 4, the reactivity of druggable cysteines in the protein kinase family of therapeutic targets were predicted based on their acidities (i.e., computed pK_a's) using a variety of rigorous pK_a calculation methods. The pK_a's of important oncogenic mutants of these targetable kinases were also computed.

1.3 Computer Modelling in Drug Discovery

Computer modelling is used extensively in the pharmaceutical industry to aid in the development of new drugs. The membrane permeability of a drug can be estimated by empirical computational methods or molecular simulation.[46–48] Docking algorithms are used to rapidly screen large databases of compounds for their ability to bind a protein or nucleic acid that is targeted for inhibition.[49–51] Other methods, such as free energy perturbation (FEP), are used to calculate the binding affinities of a drug to a protein ($\Delta G_{non-covalent}$).[52–55] These methods are generally based on molecular mechanical force fields or other simplified representations of the protein and ligand, which typically only describe the intermolecular component of protein–ligand binding. Covalent modification inherently involves the making and breaking of chemical bonds, so these methods must be adapted to describe this mode of binding.

In Chapter 5, a multiscale approach is undertaken in an effort to describe all the steps in the covalent binding process and quantify the energetics of non-covalent and covalent aspects of the binding process. The model system is a high-resolution crystal structure of Bruton's tyrosine kinase (BTK) complexed with a t-butyl cyanoacrylamide ligand bearing a piperidine linker and pyrazolopyrimidine scaffold (PDB ID: 4YHF), Figure 1.6.

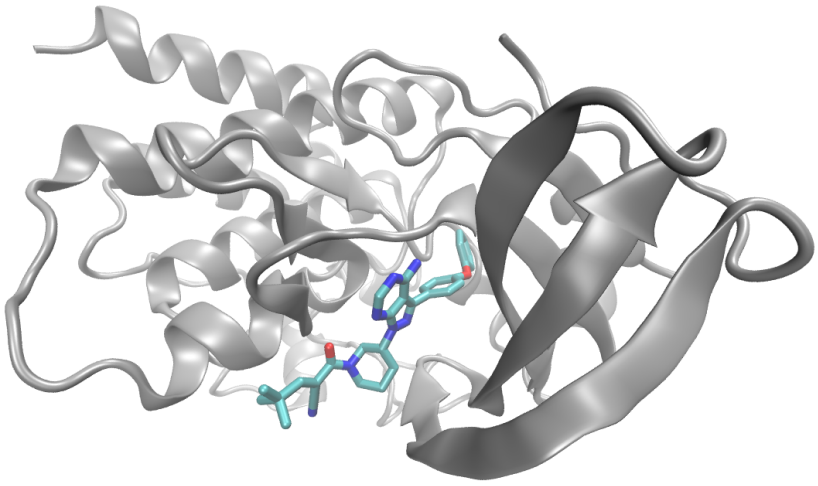

Figure 1.6: X-ray crystallographic structure of BTK complexed with t-butyl cyanoacrylamide ligand (PDB ID: 4YHF).

1.3.1 Docking and Free Energy Calculations

A principal goal of modern structure-based and computational drug design is to be able to accurately predict the binding mode, selectivity, and affinity of a potential drug candidate to a protein target. Docking algorithms and free energy calculations offer an avenue to estimate the bound-conformation and binding free energies of active molecules in their targets for lead optimization.[56]

In molecular docking, large libraries of compounds are screened for their ability to bind to a receptor. The compounds screened are then ranked based on a scoring function that takes into account the conformation and interaction energies of the compounds in the binding pocket of the receptor. In order to screen thousands of compounds, docking algorithms use highly simplified and efficient scoring functions to predict both ligand orientation and interaction, limiting their ability to quantitatively predict accurate binding strengths. Further, these methods in most cases fail to account for the conformational changes of the receptor, configurational entropy of the ligand upon binding, and solvent effects. Despite these limitations, the recent advancements in computer technology, algorithm, and speed has led to the development of more sophisticated and improved computational docking methods.[49]

Free energy calculations on the other hand, offer a more rigorous approach to calculate protein–ligand binding energies.[57] These calculations, despite being computationally more expensive than docking methods, account for conformational changes and entropy of the receptor and ligand, while also taking into account the discrete nature of the solvent as explicit water molecules. The binding free energy can be estimated from binding and unbinding events that yield an accurate thermodynamic average, so long as adequate sampling is achieved. Free energy calculation methods can be used to compute either relative binding affinities[55,58] (i.e., the difference in the binding affinity between two or more related ligands) or absolute binding affinities[19,52] (i.e., the binding affinity of a single ligand to a receptor).

For the work discussed in Chapter 5 of this thesis, the absolute binding energy of the t-butyl cyanoacrylamide ligand to BTK receptor (which represents $\Delta G_{non-covalent}$) is computed using alchemical free energy calculations. Alchemical free energy calculations[54,59,60] offer a theoretically rigorous way of computing ligand binding free energies. In alchemical free energy calculations, the ligand is slowly decoupled from its binding environment into a non-interacting "ghost" molecule in a series of intermediate

stages that characterize binding/unbinding processes. This alchemical approach of computing ligand–receptor binding free energy is computationally more tractable and cost-effective than direct sampling of the bound and unbound states of the ligand. Also, given that free energy is a state function (i.e., path independent), alchemical simulations that provide a convenient pathway connecting the final bound and unbound thermodynamic states are an efficient way of computing absolute binding free energies of druggable molecules. In fact, alchemical free energy calculations have been shown to provide accurate estimates of the binding affinity,[61] selectivity,[62–64] and specificity[65,66] of drug-like molecules binding to biologically relevant enzyme targets.

The approach for computing the absolute binding free energy of a ligand to a receptor follows a non-physical thermodynamic cycle depicted in Figure 1.7, where the binding free energy is computed through a series of alchemical transformations that characterize binding/unbinding processes of the ligand in the bound/unbound states. The absolute free energy of ligand binding ($\Delta G^{o}_{binding}$ which is synonymous to $\Delta G_{non-covalent}$) is determined from the sum of the separate energy contributions and corresponds to the forces and intermolecular interactions of the ligand following its association and dissociation from the protein. These calculations are performed for both the ligand in bulk solvent and in the protein binding site. For the example shown in Figure 1.7, the binding free energy of transferring a ligand from bulk solvent to the protein binding site consists of computing the terms: $\Delta G^{solv}_{elec+vdw}$, ΔG^{solv}_{restr}, $\Delta G^{prot}_{elec+vdw}$, and ΔG^{prot}_{restr}. The free energy terms $\Delta G^{solv}_{elec+vdw}$ and $\Delta G^{prot}_{elec+vdw}$ represent the interaction energy following the dissociation of the ligand from bulk solvent and its association in the protein binding site, respectively. ΔG^{solv}_{restr} and ΔG^{prot}_{restr} correspond to the free energy cost due to the conformational, positional, and orientational restraints underlying the ligand binding process.

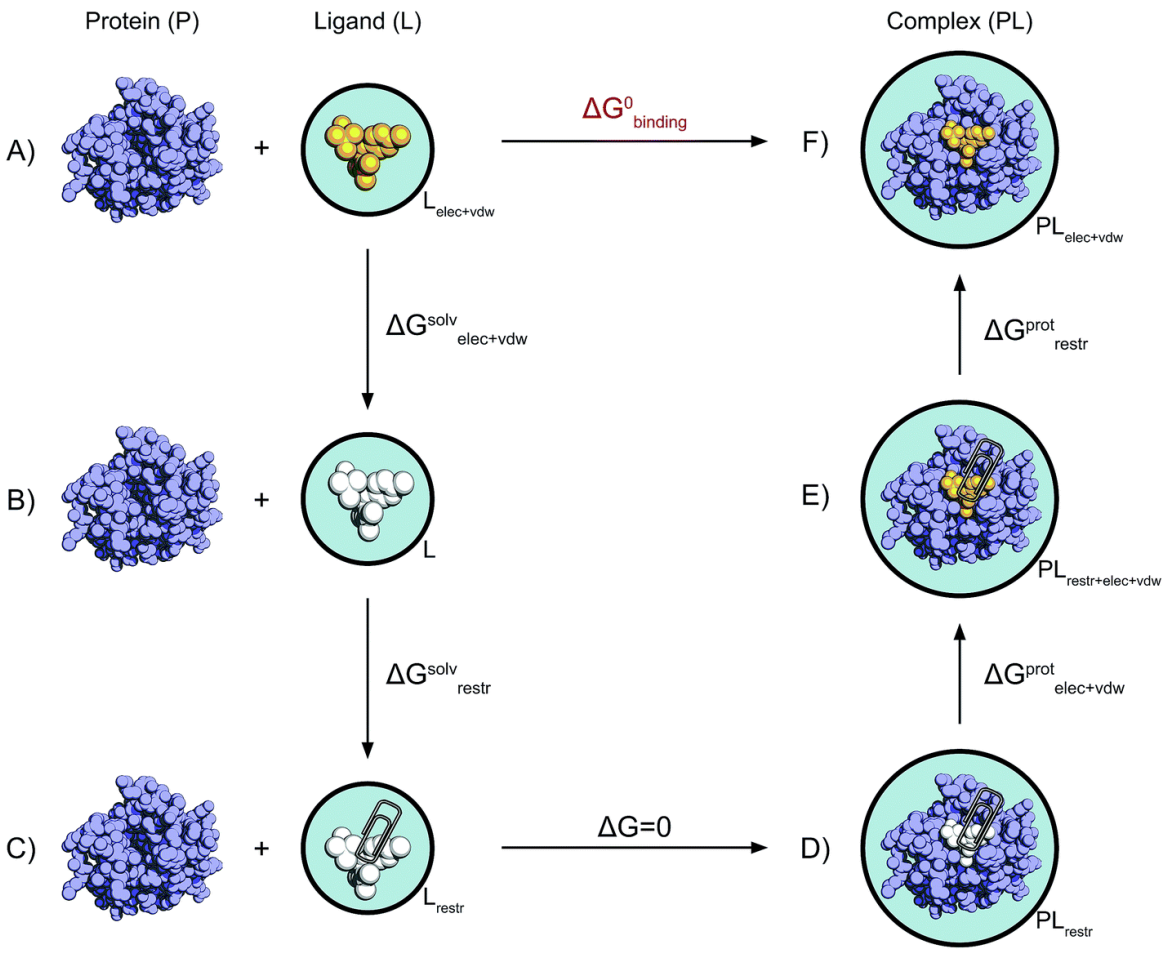

Figure 1.7: Thermodynamic cycle showing the necessary steps involved in the calculation of absolute binding free energies. The fully interacting ligand (orange) in solution at the top left (A) is transformed into a non-interacting solute (B, white) during a series of equilibrium simulations where its electrostatic and van der Waals interactions are scaled to zero, providing the term $\Delta G^{solv}_{elec+vdw}$. The ligand is then restrained while still non-interacting with the environment (C), yielding ΔG^{solv}_{restr}. This state is equivalent to having the non-interacting ligand restrained within the protein cavity (D). The restrained ligand and non-interacting ligand in complex with the protein has its electrostatic and van der Waals interactions turned back on again (E), giving $\Delta G^{prot}_{elec+vdw}$. The restraints between the ligand and protein are then removed (ΔG^{prot}_{restr}), closing the cycle, and the final state is the unrestrained and fully interacting ligand in complex with the protein (F). Adapted from Aldeghi *et al.*[61] with permission from the Royal Society of Chemistry.

1.4 Quantum Chemical Methodology

The pK_a calculation and free energy methods described thus far rely on molecular mechanical methods to describe the protein and inhibitor. Typically, these methods do not describe bond formation and breaking processes associated with chemical reactions, so terms like $\Delta G_{covalent}$ and $\Delta G^{\ddagger}$ cannot be calculated by these methods. This has led researchers to employ quantum chemistry to model the mechanisms, kinetics, and structures involved in covalent modification. Density functional theory (DFT) is widely used for modelling biological systems because of its ability to describe large chemical systems with quantitatively accurate energies and structures.[67]

Early models of electrophilic thiol additions were unable to identify the enolate/carbanion intermediates that occur in the canonical mechanism for a thiol-Michael addition.[68] The failure of conventional DFT methods to describe these reactions stems from an issue in contemporary DFT known as delocalization error.[69–72] DFT calculates inter-electron repulsion in a way that erroneously includes repulsion between an electron and itself, which must be corrected for in an approximate way through the exchange-correlation functional. The result of this effect is a spurious delocalization of electrons to reduce their self-interaction (Figure 1.8).

Delocalization error is an issue when DFT is used to model thiol additions.[68,73] The thiolate intermediate features a diffuse, sulfur-centered anion. When some popular DFT functionals are used (e.g., B3LYP or PBE), self-interaction error[74,75] causes the energy level of the highest occupied molecular orbital (HOMO) to be positive, making the anionic electron formally unbound. When the thiolate is complexed with a Michael acceptor, delocalization error spuriously stabilizes a non-bonded state where electron density is transferred from the HOMO of the thiolate to orbitals of the Michael acceptor. For some electrophiles, this complex is the most stable form and these methods predict that there is no enolate/carbanion intermediate.

Issues with delocalization error have led to the development of range-separated DFT functionals, where the exchange-correlation functional uses a large component of exact exchange for long-range inter-electron exchange-correlation. Smith et al. showed that range separated DFT functionals such as ωB97X-D predicted a stable thiocarboanion intermediate, while popular methods like B3LYP predicted that this

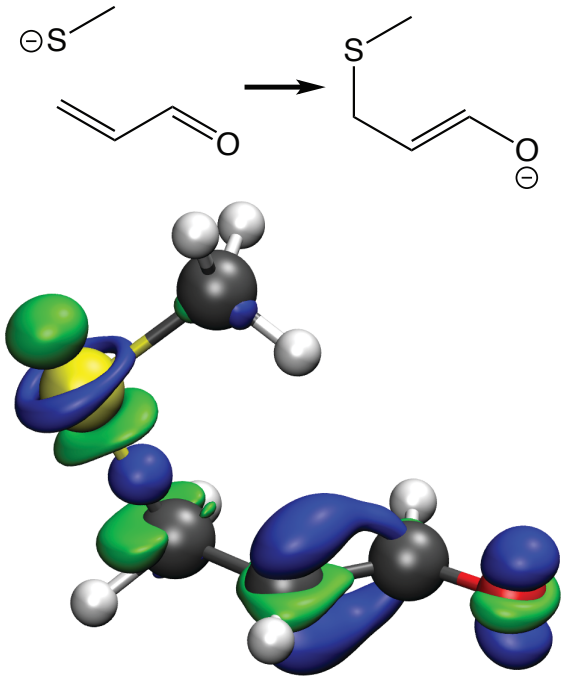

Figure 1.8: The charge transfer between a thiolate and Michael acceptor calculated using ωB97X-D/aug-cc-pVTZ. Charge is transferred from methylthiolate (top) to the acrolein Michael acceptor (bottom). Areas in blue indicate an increase in charge density while areas in green correspond to a decrease in charge density when the two fragments interact. Charge is lost from the thiolate anion and gained in the space between the S–C_α sigma bond, the π molecular orbital of the C_α–C bond, and the p_z orbital of the O atom, corresponding to an oxygen-centered anion.

intermediate could not exist as a distinct species (Figure 1.9).[68] This result was corroborated by highly accurate CCSD(T) calculations. Some hybrid functionals that have a high component of exact exchange globally, such as PBE0 or M06-2X, also predicted a stable carbanion intermediate. In this thesis, the ωB97X-D is the preferred functional of choice for modelling the chemical step of the thiol-Micheal addition reaction investigated in Chapter 5.

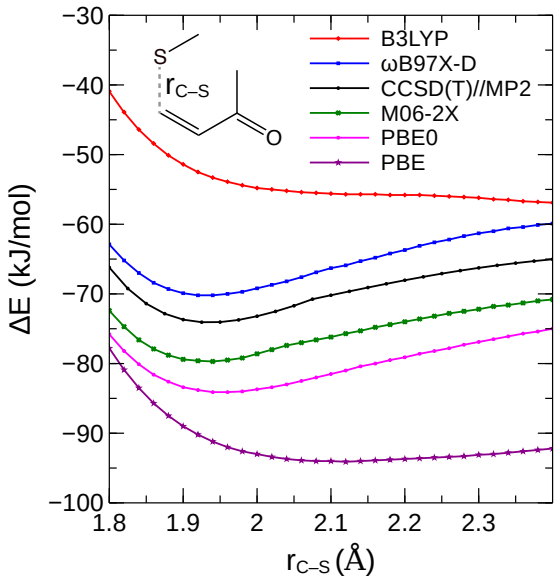

Figure 1.9: The potential energy surfaces for the addition of methylthiolate to methyl vinyl ketone, calculated using DFT and ab initio methods. The PES for the B3LYP and PBE functionals fail to predict a stable enolate intermediate. High-level ab initio (CCSD(T)), range-separated functions (e.g., ωB97X-D) and hybrid functionals (e.g., PBE0) predict a moderately-stable enolate intermediate with a minimum near C_β–S = 1.9 Å.

1.4.1 QM/MM Models of Covalent Modification

Studies of covalent modification using model reactants in the gas phase or using a continuum solvent model do not provide a rigorous description of how the protein environment affects the reaction between the protein and the inhibitor. Paasche et al. found that continuum solvent models provided limited success in describing the cysteine–histidine proton transfer reactions associated with cysteine protease function.[76] Describing the full enzyme, inhibitor, and solvent using a quantum mechanical model would be prohibitively computationally demanding, so it is not practical to apply these methods naively to model the covalent modification of a protein.

Quantum mechanics/molecular mechanics (QM/MM) methods allow for a critical component of a chemical system to be described using a quantum mechanical model, while the rest of the system is represented using a molecular mechanical model (Figure 1.10). As the size of the QM region is reduced to a relatively small size, the computational expense of these QM/MM calculations is tractable. This is well-suited

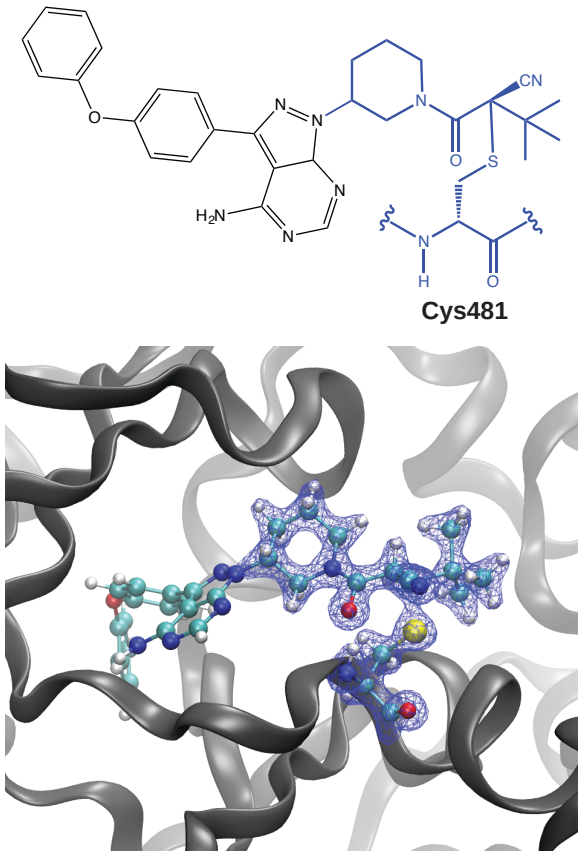

Figure 1.10: An example QM/MM model of Bruton's tyrosine kinase in complex with a covalent modifier (Ref. 23, PDB ID: 4YHF). The covalently-modified Cys481 residue and the cyanoacrylamide warhead define the QM region. The calculated electron density of the QM region is represented by the blue mesh. The remainder of the protein (gray) and inhibitor comprise the MM region.

for modelling chemical reactions involving proteins, like enzymatic reaction mechanisms, where the chemical reaction only directly involves a small number of atoms, but the rest of the system provides an essential environment. Analogously, the covalent modification of proteins can also be described using a QM/MM model, where the reactive warhead of the inhibitor and the residue being modified are described using QM, while the balance of the system, such as the solvent and the rest of the protein are described using an MM model. If needed, additional sections of the inhibitor and protein can be included in the QM region.

QM/MM modelling has the potential to play a significant role in understanding

and predicting the mechanisms, kinetics, and thermodynamics of covalent modification. In this thesis, QM/MM method is used to model the covalent modification steps of a cyanoacrylamide inhibitor binding to BTK and to calculate a rigorous, complete binding energy profile of the chemical reaction (Chapter 5). This approach allows us to model the action of covalent modifier drugs in a comprehensive way through the calculation of $\Delta G_{non-covalent}$, $\Delta G^{\ddagger}$, and $\Delta G_{covalent}$. We also used QM/MM to study the solvation structure of model thiolates in aqueous solutions in an effort to determine their hydration structure (Chapter 3).

1.5 Outline

In this thesis, I explore the reactivity of cysteine residues in proteins from a computational chemistry standpoint. Methods to identify which cysteines in druggable protein targets that have the right balance between reactivity and *in vivo* stability inform drug development.

In Chapter 2, I perform benchmark assessments of different computational methods in accurately estimating experimental cysteine pK_a's for a test set of proteins. Methods that use either an implicit solvent or explicit solvent model were analyzed to determine their predictive accuracy. In an effort to address the intrinsic limitations in the accuracy of pK_a calculation methods as observed in chapter 2, Chapter 3 details an approach to determine which molecular mechanical model provides the best description for model thiolates in solution. The cysteine thiolate parameters for the Amber and CHARMM force field models are validated using free energy perturbation methods and QM/MM MD simulations. In Chapter 4, improved pK_a calculation methods and rigorous computational approaches like constant-pH molecular dynamics that are capable of describing variable protonation states within proteins, are used to predict the reactivity of select druggable cysteines across the protein kinase family of popular drug targets. Important oncogenic mutants of these kinases are also included in the test set. Chapter 5 presents the first complete model for covalent modification of druggable cysteine in an enzyme target, including both noncovalent and covalent binding steps. The absolute binding free energy is calculated rigorously using advanced MD and hybrid QM/MM methods. A brief summary of the research results is presented in Chapter 6, along with future directions stemming from this work.

Bibliography

[1] Kuntz, I. D.; Chen, K.; Sharp, K. A.; Kollman, P. A. The Maximal Affinity of Ligands. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 9997–10002.

[2] Jackson, P. A.; Widen, J. C.; Harki, D. A.; Brummond, K. M. Covalent Modifiers: A Chemical Perspective on the Reactivity of α,β-Unsaturated Carbonyls with Thiols via Hetero-Michael Addition Reactions. *J. Med. Chem.* **2017**, *60*, 839–885.

[3] Potashman, M. H.; Duggan, M. E. Covalent Modifiers: An Orthogonal Approach to Drug Design. *J. Med. Chem.* **2009**, *52*, 1231–1246.

[4] Singh, J.; Petter, R. C.; Baillie, T. A.; Whitty, A. The resurgence of covalent drugs. *Nat. Rev. Drug Discov.* **2011**, *10*, 307–317.

[5] Kalgutkar, A. S.; Dalvie, D. K. Drug discovery for a new generation of covalent drugs. *Expert Opin. Drug Discov.* **2012**, *7*, 561–581.

[6] Wilson, A. J.; Kerns, J. K.; Callahan, J. F.; Moody, C. J. Keap Calm, and Carry on Covalently. *J. Med. Chem.* **2013**, *56*, 7463–7476.

[7] Bauer, R. A. Covalent inhibitors in drug discovery: from accidental discoveries to avoided liabilities and designed therapies. *Drug Discov. Today* **2015**, *20*, 1061–1073.

[8] Baillie, T. A. Targeted Covalent Inhibitors for Drug Design. *Angew. Chem. Int. Ed.* **2016**, *55*, 13408–13421.

[9] Sotriffer, C. Docking of Covalent Ligands: Challenges and Approaches. *Mol. Inf.* **2018**, *37*, 1800062.

[10] Ghosh, A. K.; Samanta, I.; Mondal, A.; Liu, W. R. Covalent Inhibition in Drug Discovery. *ChemMedChem* **2019**, *14*, 889–906.

[11] Powers, J. C.; Asgian, J. L.; Ekici, Ö. D.; James, K. E. Irreversible Inhibitors of Serine, Cysteine, and Threonine Proteases. *Chem. Rev.* **2002**, *102*, 4639–4750.

[12] Cuesta, A.; Taunton, J. Lysine-Targeted Inhibitors and Chemoproteomic Probes. *Annu. Rev. Biochem.* **2019**, *88*, 365–381.

[13] Miseta, A.; Csutora, P. Relationship between the occurrence of cysteine in proteins and the complexity of organisms. *Mol. Biol. Evol.* **2000**, *17*, 1232–1239.

[14] Jöst, C.; Nitsche, C.; Scholz, T.; Roux, L.; Klein, C. D. Promiscuity and Selectivity in Covalent Enzyme Inhibition: A Systematic Study of Electrophilic Fragments. *J. Med. Chem.* **2014**, *57*, 7590–7599.

[15] Backus, K. M.; Correia, B. E.; Lum, K. M.; Forli, S.; Horning, B. D.; González-Páez, G. E.; Chatterjee, S.; Lanning, B. R.; Teijaro, J. R.; Olson, A. J.; Wolan, D. W.; Cravatt, B. F. Proteome-wide covalent ligand discovery in native biological systems. *Nature* **2016**, *534*, 570–574.

[16] Parker, C. G.; Galmozzi, A.; Wang, Y.; Correia, B. E.; Sasaki, K.; Joslyn, C. M.; Kim, A. S.; Cavallaro, C. L.; Lawrence, R. M.; Johnson, S. R.; Narvaiza, I.; Saez, E.; Cravatt, B. F. Ligand and Target Discovery by Fragment-Based Screening in Human Cells. *Cell* **2017**, *168*, 527–541.e29.

[17] Johnson, D. S.; Weerapana, E.; Cravatt, B. F. Strategies for discovering and derisking covalent, irreversible enzyme inhibitors. *Future Med. Chem.* **2010**, *2*, 949–964.

[18] Johansson, M. H. Reversible Michael Additions: Covalent Inhibitors and Prodrugs. *Mini-Rev. Med. Chem.* **2012**, *12*, 1330–1344.

[19] Woo, H.-J.; Roux, B. Calculation of absolute protein–ligand binding free energy from computer simulations. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6825–6830.

[20] Nagel, Z. D.; Klinman, J. P. A 21st century revisionist's view at a turning point in enzymology. *Nat. Chem. Biol.* **2009**, *5*, 543–550.

[21] Olsson, M. H.; Mavri, J.; Warshel, A. Transition state theory can be used in studies of enzyme catalysis: lessons from simulations of tunnelling and dynamical effects in lipoxygenase and other systems. *Philos. Trans. R. Soc., B* **2006**, *361*, 1417–1432.

[22] Copeland, R. A.; Pompliano, D. L.; Meek, T. D. Drug-target residence time and its implications for lead optimization. *Nat. Rev. Drug Discov.* **2006**, *5*, 730–739.

[23] Bradshaw, J. M. et al. Prolonged and tunable residence time using reversible covalent kinase inhibitors. *Nat. Chem. Biol.* **2015**, *11*, 525–531.

[24] Copeland, R. A. The drug-target residence time model: a 10-year retrospective. *Nat. Rev. Drug Discov.* **2016**, *15*, 87–95, Perspectives.

[25] Laurents, D. V.; Huyghues-Despointes, B. M. P.; Bruix, M.; Thurlkill, R. L.; Schell, D.; Newsom, S.; Grimsley, G. R.; Shaw, K. L.; Treviño, S.; Rico, M.; Briggs, J. M.; Antosiewicz, J. M.; Scholtz, J. M.; Pace, C. N. Charge-charge interactions are key determinants of the pK values of ionizable groups in ribonuclease Sa (pI=3.5) and a basic variant (pI=10.2). *J. Mol. Biol.* **2003**, *325*, 1077–1092.

[26] Pace, C. N.; Grimsley, G. R.; Scholtz, J. M. Protein Ionizable Groups: pK Values and Their Contribution to Protein Stability and Solubility. *J. Biol. Chem.* **2009**, *284*, 13285–13289.

[27] Thurlkill, R. L.; Grimsley, G. R.; Scholtz, J. M.; Pace, C. N. pK Values of the Ionizable Groups of Proteins. *Protein Sci.* **2006**, *15*, 1214–1218.

[28] Pinitglang, S.; Watts, A. B.; Patel, M.; Reid, J. D.; Noble, M. A.; Gul, S.; Bokth, A.; Naeem, A.; Patel, H.; Thomas, E. W.; Sreedharan, S. K.; Verma, C.; Brocklehurst, K. A Classical Enzyme Active Center Motif Lacks Catalytic Competence Until Modulated Electrostatically. *Biochemistry* **1997**, *36*, 9968–9982.

[29] Bulaj, G.; Kortemme, T.; Goldenberg, D. P. Ionization–Reactivity Relationships for Cysteine Thiols in Polypeptides. *Biochemistry* **1998**, *37*, 8965–8972.

[30] Tolbert, B. S.; Tajc, S. G.; Webb, H.; Snyder, J.; Nielsen, J. E.; Miller, B. L.; Basavappa, R. The Active Site Cysteine of Ubiquitin-Conjugating Enzymes Has a Significantly Elevated pKa: Functional Implications. *Biochemistry* **2005**, *44*, 16385–16391.

[31] Takano, K.; Scholtz, J. M.; Sacchettini, J. C.; Pace, C. N. The contribution of polar group burial to protein stability is strongly context-dependent. *J. Biol. Chem.* **2003**, *278*, 31790–31795.

[32] Reijenga, J.; van Hoof, A.; van Loon, A.; Teunissen, B. Development of methods for the determination of pKa values. *Anal. Chem. Insights* **2013**, *8*, 53–71.

[33] Combining Conformational Flexibility and Continuum Electrostatics for Calculating pKas in Proteins. *Biophys. J.* **2002**, *83*, 1731–1748.

[34] Li, G.; Cui, Q. pKa Calculations with QM/MM Free Energy Perturbations. *J. Phys. Chem. B* **2003**, *107*, 14521–14528.

[35] Kuhn, B.; Kollman, P. A.; Stahl, M. Prediction of pKa Shifts in Proteins Using a Combination of Molecular Mechanical and Continuum Solvent Calculations. *J. Comput. Chem.* **2004**, *25*, 1865–1872.

[36] Gordon, J. C.; Myers, J. B.; Folta, T.; Shoja, V.; Heath, L. S.; Onufriev, A. H++: a server for estimating pKas and adding missing hydrogens to macromolecules. *Nucleic Acids Res.* **2005**, *33*, W368–W371.

[37] Jensen, J. H.; Li, H.; Robertson, A. D.; Molina, P. A. Prediction and Rationalization of Protein pKa Values Using QM and QM/MM Methods. *J. Phys. Chem. A* **2005**, *109*, 6634–6643.

[38] Li, H.; Robertson, A. D.; Jensen, J. H. Very fast empirical prediction and rationalization of protein pKa values. *Proteins: Struct., Funct., Bioinf.* **2005**, *61*, 704–721.

[39] Riccardi, D.; Schaefer, P.; Cui, Q. pKa Calculations in Solution and Proteins with QM/MM Free Energy Perturbation Simulations: A Quantitative Test of QM/MM Protocols. *J. Phys. Chem. B* **2005**, *109*, 17715–17733.

[40] Spassov, V. Z.; Yan, L. A Fast and Accurate Computational Approach to Protein Ionization. *Protein Sci.* **2008**, *17*, 1955–1970.

[41] Huang, R.-B.; Du, Q.-S.; Wang, C.-H.; Liao, S.-M.; Chou, K.-C. A Fast and Accurate Method for Predicting pKa of Residues in Proteins. *Protein Eng. Des. Sel.* **2010**, *23*, 35–42.

[42] Alexov, E.; Mehler, E. L.; Baker, N.; M. Baptista, A.; Huang, Y.; Milletti, F.; Erik Nielsen, J.; Farrell, D.; Carstensen, T.; Olsson, M. H. M.; Shen, J. K.; Warwicker, J.; Williams, S.; Word, J. M. Progress in the Prediction of pKa Values in Proteins. *Proteins: Struct., Funct., Bioinf.* **2011**, *79*, 3260–3275.

[43] Bashford, D.; Karplus, M. pKa's of Ionizable Groups in Proteins: Atomic Detail From a Continuum Electrostatic Model. *Biochemistry* **1990**, *29*, 10219–10225.

[44] Simonson, T.; Carlsson, J.; Case, D. A. Proton Binding to Proteins: pKa Calculations with Explicit and Implicit Solvent Models. *J. Am. Chem. Soc.* **2004**, *126*, 4167–4180.

[45] Olsson, M. H. M.; Søndergaard, C. R.; Rostkowski, M.; Jensen, J. H. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions. *J. Chem. Theory Comput.* **2011**, *7*, 525–537.

[46] Mannhold, R.; Poda, G. I.; Ostermann, C.; Tetko, I. V. Calculation of Molecular Lipophilicity: State-of-the-art and comparison of logP methods on more than 96,000 compounds. *J. Pharm. Sci.* **2009**, *98*, 861–893.

[47] Awoonor-Williams, E.; Rowley, C. N. Molecular simulation of nonfacilitated membrane permeation. *Biochim. Biophys. Acta, Biomembr.* **2016**, *1858*, 1672–1687.

[48] Lee, C. T.; Comer, J.; Herndon, C.; Leung, N.; Pavlova, A.; Swift, R. V.; Tung, C.; Rowley, C. N.; Amaro, R. E.; Chipot, C.; Wang, Y.; Gumbart, J. C. Simulation-Based Approaches for Determining Membrane Permeability of Small Compounds. *J. Chem. Inf. Model.* **2016**, *56*, 721–733.

[49] Yuriev, E.; Agostino, M.; Ramsland, P. A. Challenges and advances in computational docking: 2009 in review. *J. Mol. Recogn.* **2011**, *24*, 149–164.

[50] Cheng, T.; Li, Q.; Zhou, Z.; Wang, Y.; Bryant, S. H. Structure-Based Virtual Screening for Drug Discovery: a Problem-Centric Review. *AAPS J.* **2012**, *14*, 133–141.

[51] Sliwoski, G.; Kothiwale, S.; Meiler, J.; Lowe, E. W. Computational Methods in Drug Discovery. *Pharmacol. Rev.* **2013**, *66*, 334–395.

[52] Wang, J.; Deng, Y.; Roux, B. Absolute Binding Free Energy Calculations Using Molecular Dynamics Simulations with Restraining Potentials. *Biophys J.* **2006**, *91*, 2798–2814.

[53] Michel, J.; Essex, J. W. Prediction of protein–ligand binding affinity by free energy simulations: assumptions, pitfalls and expectations. *J. Comput. Aided Mol. Des.* **2010**, *24*, 639–658.

[54] Chodera, J. D.; Mobley, D. L.; Shirts, M. R.; Dixon, R. W.; Branson, K.; Pande, V. S. Alchemical free energy methods for drug discovery: progress and challenges. *Curr. Opin. Struct. Biol.* **2011**, *21*, 150–160.

[55] Wang, L. et al. 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*, 2695–2703.

[56] Gilson, M. K.; Zhou, H.-X. Calculation of protein-ligand binding affinities. *Annu. Rev. Biophys. Biomol. Struct.* **2007**, *36*, 21–42.

[57] de Ruiter, A.; Oostenbrink, C. Free energy calculations of protein–ligand interactions. *Curr. Opin. Chem. Biol.* **2011**, *15*, 547–552.

[58] Cournia, Z.; Allen, B.; Sherman, W. Relative binding free energy calculations in drug discovery: recent advances and practical considerations. *J. Chem. Inf. Model.* **2017**, *57*, 2911–2937.

[59] Shirts, M. R.; Mobley, D. L.; Chodera, J. D. Alchemical free energy calculations: ready for prime time? *Annu Rep Comput Chem* **2007**, *3*, 41–59.

[60] Mobley, D. L.; Klimovich, P. V. Perspective: Alchemical free energy calculations for drug discovery. *J. Chem. Phys.* **2012**, *137*, 230901.

[61] Aldeghi, M.; Heifetz, A.; Bodkin, M. J.; Knapp, S.; Biggin, P. C. Accurate calculation of the absolute free energy of binding for drug molecules. *Chem. Sci.* **2016**, *7*, 207–218.

[62] Lin, Y.-L.; Meng, Y.; Huang, L.; Roux, B. Computational study of Gleevec and G6G reveals molecular determinants of kinase inhibitor selectivity. *J. Am. Chem. Soc.* **2014**, *136*, 14753–14762.

[63] Alsamarah, A.; LaCuran, A. E.; Oelschlaeger, P.; Hao, J.; Luo, Y. Uncovering molecular bases underlying bone morphogenetic protein receptor inhibitor selectivity. *PloS One* **2015**, *10*.

[64] Aldeghi, M.; Heifetz, A.; Bodkin, M. J.; Knapp, S.; Biggin, P. C. Predictions of ligand selectivity from absolute binding free energy calculations. *J. Am. Chem. Soc.* **2017**, *139*, 946–957.

[65] Lin, Y.-L.; Meng, Y.; Jiang, W.; Roux, B. Explaining why Gleevec is a specific and potent inhibitor of Abl kinase. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110*, 1664–1669.

[66] Lin, Y.-L.; Roux, B. Computational analysis of the binding specificity of Gleevec to Abl, c-Kit, Lck, and c-Src tyrosine kinases. *J. Am. Chem. Soc.* **2013**, *135*, 14741–14753.

[67] Jones, R. O. Density functional theory: Its origins, rise to prominence, and future. *Rev. Mod. Phys.* **2015**, *87*, 897.

[68] Smith, J. M.; Jami Alahmadi, Y.; Rowley, C. N. Range-separated DFT functionals are necessary to model Thio-Michael additions. *J. Chem. Theory Comput.* **2013**, *9*, 4860–4865.

[69] Johnson, E. R.; Mori-Sánchez, P.; Cohen, A. J.; Yang, W. Delocalization errors in density functionals and implications for main-group thermochemistry. *J. Chem. Phys.* **2008**, *129*, 204112.

[70] Cohen, A. J.; Mori-Sánchez, P.; Yang, W. Insights into Current Limitations of Density Functional Theory. *Science* **2008**, *321*, 792–794.

[71] Autschbach, J.; Srebro, M. Delocalization Error and "Functional Tuning" in Kohn–Sham Calculations of Molecular Properties. *Acc. Chem. Res.* **2014**, *47*, 2592–2602.

[72] Wasserman, A.; Nafziger, J.; Jiang, K.; Kim, M.-C.; Sim, E.; Burke, K. The Importance of Being Self-Consistent. *Annu. Rev. Phys. Chem.* **2017**,

[73] Awoonor-Williams, E.; Isley III, W. C.; Dale, S. G.; Johnson, E. R.; Yu, H.; Becke, A. D.; Roux, B.; Rowley, C. N. Quantum Chemical Methods for Modeling Covalent Modification of Biological Thiols. *J. Comput. Chem.* **2020**, *41*, 427–438.

[74] Lundberg, M.; Siegbahn, P. E. Quantifying the effects of the self-interaction error in DFT: When do the delocalized states appear? *J. Chem. Phys.* **2005**, *122*, 224103.

[75] Bao, J. L.; Gagliardi, L.; Truhlar, D. G. Self-interaction error in density functional theory: An appraisal. *J. Phys. Chem. Lett.* **2018**, *9*, 2353–2358.

[76] Paasche, A.; Schirmeister, T.; Engels, B. Benchmark Study for the Cysteine–Histidine Proton Transfer Reaction in a Protein Environment: Gas Phase, COSMO, QM/MM Approaches. *J. Chem. Theory Comput.* **2013**, *9*, 1765–1777.

"We can only see a short distance ahead, but we can see plenty there that needs to be done."

— Alan Turing

2

Evaluation of Methods for the Calculation of Cysteine pK_a in Proteins

This chapter is adapted with permission from: Awoonor-Williams, E. and Rowley, C. N. Evaluation of Methods for the Calculation of the pK_a of Cysteine Residues in Proteins *J. Chem. Theory Comput.*, **2016**, 12 (9), 4662–4673.

Contents

2.1 Abstract

Methods for the calculation of the pK_a of ionizable amino acids are valuable tools for understanding pH-dependent properties of proteins. Cysteine is unique among the amino acids because of the chemical reactivity of its thiol group (S–H), which plays an instrumental role in several biochemical and regulatory functions. The acidity of noncatalytic cysteine residues is a factor in their susceptibility to chemical modification. Despite the plethora of existing pK_a computing methods, no definitive protocol exists for accurately calculating the pK_a's of cysteine residues in proteins. A cysteine pK_a test set was developed, which is comprised of 18 cysteine residues in 12 proteins where the pK_a's have been determined experimentally and an experimental structure is available. The pK_a's of these residues were calculated using three methods that use an implicit solvent model (H++, MCCE, and PROPKA) and an all-atom replica exchange thermodynamic integration approach with the CHARMM36 and AMBER ff99SB-ILDNP force fields. The models that use implicit solvation methods were generally unreliable in predicting cysteine residue pK_a's, with RMSDs between 3.41 and 4.72 pK_a units. On average, the explicit solvent methods performed better than the implicit solvent methods. RMSD values of 2.40 and 3.20 were obtained for simulations with the CHARMM36 and AMBER ff99SB-ILDNP force fields, respectively. Further development of these methods is necessary because the performance of the best method is similar to that of the null-model (RMSD=2.74) and these differences in RMSD are of limited statistical significance given the small size of our test set.

2.2 Introduction

Cysteine is unique among the amino acids because of its thiol (S–H) functional group. This moiety allows cysteine to serve several biochemical roles,[1,2] including disulfide bond formation,[3] metal-binding,[4] proton donor,[5,6] and redox-catalyst.[6,7] Many of these processes require cysteine to act as Brønsted–Lowry acid. The relative weakness of the S–H bond allows these reactions to occur spontaneously under mild conditions; cysteine has an intrinsic pK_a of ∼8.6,[8] which is one of the closest to physiological pH of all the naturally occurring amino acids.

In many instances, the reaction mechanism of the cysteine residue involves the

formation of a negatively-charged thiolate anion. The catalytic cycles of the cysteine protease[5,9] and protein tyrosine phosphatase[10,11] enzyme families involve the deprotonation of a cysteine residue as a critical step. For instance, in the cysteine protease family, the thiolate is a necessary intermediate that undergoes a nucleophilic attack on the carbonyl carbon of an amide bond (Figure 2.1 (a)).

Cysteine side chains in proteins also engage in a broad range of chemical reactions with both endogenous and exogenous compounds. For example, the anti-cancer drugs neratinib, dacomitinib, and afatinib contain an electrophilic acrylamide moiety that inhibits target kinase enzymes by undergoing addition to an active site cysteine.[12–15] The canonical mechanism for these reactions is a thiol-Michael addition, where the thiol side-chain of the cysteine must be deprotonated before reacting with an electrophilic carbon of the acrylamide moiety (Figure 2.1 (b)). The covalent modification of the epidermal growth factor receptor by afatinib is an example of this activity (Figure 2.2).

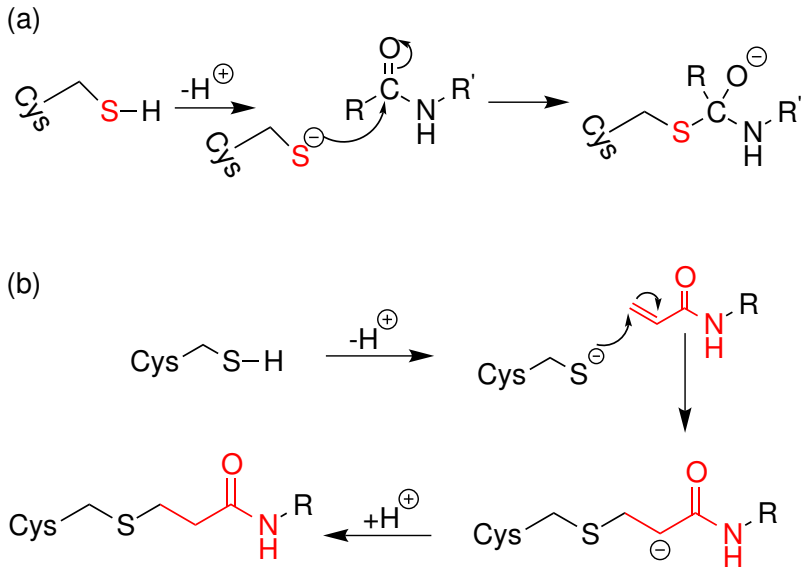

Figure 2.1: Mechanism of cysteine reactivity (a) in cysteine protease enzymes and (b) with the acrylamide moiety (red) of a covalent modifier drug.

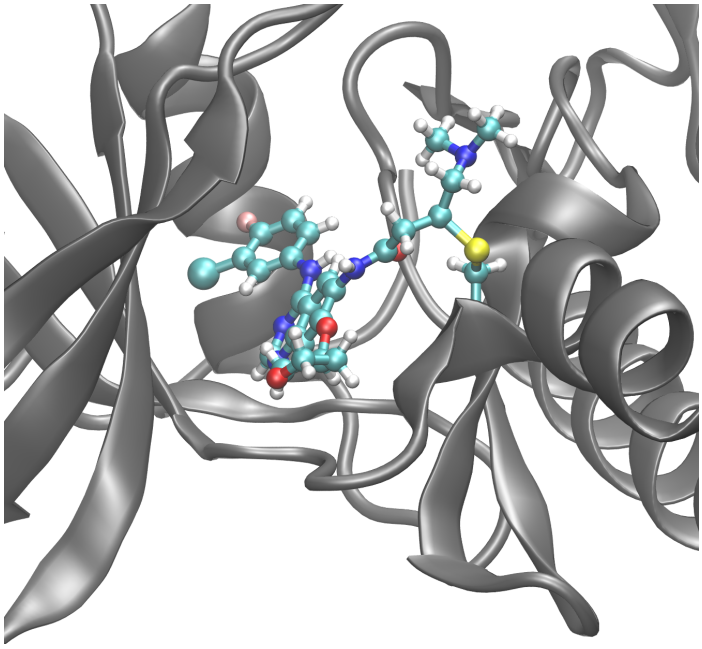

Figure 2.2: Covalent-modifier afatinib bound to the kinase domain of the epidermal growth factor receptor (EGFR). The acrylamide moiety of afatinib has formed a covalent bond with the protein by undergoing a Michael addition with the thiol group of Cys797.

2.2.1 Factors Affecting Cysteine pK_a's

The pK_a of a cysteine residue in a protein can be shifted significantly from its intrinsic pK_a value. Catalytic cysteines have pK_a's that are as low as 2.88,[16–18] but the pK_a's of non-catalytic cysteines generally range from 7.4 to 9.1.[19] Generally, intermolecular interactions that stabilize the thiolate or destabilize the thiol state will lower the pK_a, while intermolecular interactions that destabilize the thiolate and stabilize the thiol will raise the pK_a. The thiol group is a poor hydrogen bonding partner, but the thiolate group is a good hydrogen bond acceptor, so a hydrogen bond donor near a cysteine residue can decrease its pK_a.[19] Electrostatic interactions between the thiolate and charged residues is another important determinant.[20–23] Positively charged residues will have an attractive interaction with the thiolate that will lower the pK_a, while negatively charged residues will have repulsive interactions with the thiolate that will raise the pK_a.

2.2.2 Methods of pK_a Determination

The pK_a's of cysteine residues in proteins are typically determined by spectroscopic observation of the thiol during pH titration[24,25] or by measuring the rate of a chemical reaction involving the thiol.[26,27] Recently, potentiometric titration[28] and quantitative mass spectrometry[29] have been explored as alternative methods for cysteine pK_a determination. These methods require the protein of interest to be expressed or synthesized with good yield and purity. Furthermore, the presence of multiple ionizable residues within the protein complicates these experiments.

The challenges of experimental pK_a determination have spurred the development of a number of computational techniques to estimate the pK_a's of amino acids in proteins.[30–38] Using the three dimensional protein structure determined by X-ray crystallography or Nuclear Magnetic Resonance (NMR) spectroscopy, these algorithms can estimate the pK_a of a given residue based on the relative stability of protonated and deprotonated forms of the side chain. These methods can be divided into those that model the effect of the solvent by an implicit representation and those that include explicit representations of solvent water molecules.

Implicit solvent methods are popular because they generally have a lower computational cost and the calculations can often be performed with limited preparation by the user. These methods generally model the solvent around the protein as a continuum dielectric.[39–41] The electrostatic effect on the pK_a of a residue is determined by numerical solutions to either the Poisson–Boltzmann Equation (PBE)[39,41–45] or by Generalized Born theory.[46,47] H++[36,48,49] and Multi-Conformation Continuum Electrostatics (MCCE)[30,50,51] are two popular pK_a prediction methods based on a continuum electrostatic solvent model.

PROPKA is a popular empirical pK_a prediction method that also describes solvation effects implicitly.[33,52–54] Given a protein structure, this method estimates pK_a's based on approximate perturbative terms, including desolvation of the residue, interresidue Coulombic interactions, and hydrogen bonding interactions. The methods used to calculate these terms are approximate, but they are parameterized based on a large set of experimental pK_a's. PROPKA is one of the most accurate pK_a prediction methods;[55,56] the pK_a's of Asp and Glu side chains are predicted with root-mean-square deviations (RMSDs) of 0.79.[53] Previous surveys have shown that this method performs well for Tyr, Lys, and His side chains as well,[53] although no study has

systematically examined its accuracy for cysteine.

There are intrinsic limitations to the accuracy of these methods. Solvation has a large effect on pK_a, so the simplifications associated with using an implicit solvent model are significant sources of error. Moreover, the PROPKA and H++ methods assume that the protein structure is static, while the MCCE method only incorporates the effect of side chain rotations. This neglects the true dynamic nature of the protein, particularly in terms of how the structure changes when the cysteine residue is deprotonated. These limitations can be addressed by modelling the protein using molecular dynamics simulation methods that represent the solvent explicitly. Free energy calculation methods like free energy perturbation and thermodynamic integration can be used to calculate the relative Gibbs energies of the protonated and deprotonated states of the protein, which in turn can be related to the relative pK_a of a cysteine residue.[57]

2.2.3 Need for Validation

The available pK_a prediction methods vary enormously in computational cost and can be more accurate for some residues than others.[56,58] Additionally, there is no definitive protocol for accurately calculating the pK_a's of cysteines.[59] Benchmark studies on the performance of common pK_a prediction methods have shown that the accuracy of these methods varies widely.[55,60] A comprehensive comparison and evaluation of existing pK_a computing methods will allow for a more thorough assessment of their accuracy and identify where improvement is needed.

In this study, we have calculated the pK_a's of 18 cysteine residues in 12 proteins where an experimental value has been reported. The pK_a's of the selected cysteine residues were calculated using three popular implicit solvation methods, namely: H++,[49] MCCE[51] and PROPKA.[54] We also compute the pK_a's of the selected cysteine residues using an all-atom replica-exchange thermodynamic integration (RETI) explicit solvent approach with the CHARMM36[61] and AMBER ff99SB-ILDNP[62] force fields.

2.3 Theory and Computational Methodology

2.3.1 Test Set

The coordinates for all proteins modeled in this study were based on structures determined using X-ray diffraction and NMR that have been deposited in the Protein Data Bank (PDB).[63] The cysteine residues in these proteins for which the pK_a's are determined do not form disulfide bonds, so their pK_a's can be accessed directly. In cases where the coordinates of some residues within the protein structure were missing, SWISS-MODEL homology modelling[64] was used to estimate the missing coordinates. The PDB identifiers of the test set structures used in our study are included in Table 2.1.

Table 2.1: Test set of protein cysteine pK_a's.

Protein Abbrev.	PDB code[a]	Cys Residue	Exptl. pK_a	Ref.[b]
α-1-AT	1QLP[65]	232	6.86 ± 0.05	[66]
ACBP-M46C	1NTI	46	8.20 ± 0.10	[67]
ACBP-S65C	1NTI	65	9.00 ± 0.10	[67]
ACBP-T17C	1NTI	17	9.80 ± 0.10	[67]
AhpC	4MA9[68]	46	5.94 ± 0.10	[69]
DJ-1	1P5F[70]	106	5.40 ± 0.10	[71]
HMCK	1I0E[72]	283	5.60 ± 0.10	[73]
HMCK-S285A	1I0E[72]	283	6.70 ± 0.10	[73]
Mb-G124C	2MGE[74]	124	6.53 ± 0.05	[75]
Mb-A125C	2MGE[74]	125	8.43 ± 0.03	[75]
MmsrA	2L90[76]	72	7.20 ± 0.20	[77]
MmsrA-E115Q	2L90[76]	72	8.20 ± 0.10	[77]
O^6-AGT	1EH6[78]	145	5.30 ± 0.20	[79]
Papain	1PPN[80]	25	3.32 ± 0.01	[16]
PTP1B	2HNP[81]	215	5.57 ± 0.12	[18]
ppΩ	1PPO[82]	25	2.88 ± 0.02	[16]
YopH	1YPT[83]	403	4.67 ± 0.15	[17]
YopH-H402A	1YPT[83]	403	7.35 ± 0.04	[17]

[a] PDB structure code, followed by the crystal structure reference number (if one was published).
[b] Reference that provides the experimental pK_a value of the Cys residue.

In the following sections, we present the relevant details about the methodology and simulation techniques we employed in calculating the pK_a's of selected cysteine

residues in the proteins. The pK_a's range between 2.88 for Papaya Protease Omega (ppΩ) to 9.80 for the T17C mutant of Acyl-coenzyme A binding protein (ACBP).

2.3.2 Implicit Solvent Methods

The pK_a's of cysteine residues in the test set were calculated using three popular pK_a prediction methods: H++, MCCE, and PROPKA. Each of these methods assigns the protonation states of residues automatically, so the calculations are based only on the "deprotonated" experimental structure without any non-standard charge assignments by the user. These methods are briefly described here:

H++

H++ computes pK_a's of titratable groups by calculating the solvation energy of the various protonation states of a protein using a continuum solvent model. The electrostatic component of the relative stability of these states is calculated by a numerical solution to the Poisson–Boltzmann equation. The accuracy of this method was evaluated on a test set of measured pK_a values for 23 proteins with 201 titratable residues (of which 81 were Cys), collected from the work of Grimsley et al.[84]. H++ predicted the correct protonation state of titratable residues 97% of the time. The true direction of pK shift was predicted 85% of the time. H++ is available through a free web interface.[85] The calculations reported here used H++ 3.1. The server default settings were used; the salt concentration was 0.15 M and the dielectric constants of the bulk protein and water environment were 10 and 80 respectively.

MCCE

MCCE provides pK_a estimates using a molecular mechanical force field with a continuum solvent model. Side chain conformations are sampled using a Monte Carlo algorithm. MCCE can predict the pK_a's of ionizable groups in proteins with an overall RMSD of 0.90 with 75% of the errors < 1 pH unit.[51] An extensive benchmark study on a large pK_a dataset of 100 proteins consisting of over 700 titratable residues, found that MCCE can predict the pK_a's of buried residues to an accuracy of 53% within 1 pK_a unit.[55] No Cys and Arg residues were included in this test set. MCCE 2.7 was

used for the calculations reported in this study. Default dielectric constants of 8 and 80 were used in pK_a calculations for the protein and the solvent, respectively.

PROPKA

PROPKA is a fast empirical method designed for structure-based pK_a prediction and rationalization of ionizable residues in proteins and protein–ligand complexes. The pK_a of a residue is estimated by calculating the pK_a shifts of titratable groups using empirical rules that incorporate effects from hydrogen bonding, desolvation, and Coulombic interactions. PROPKA can predict the pK_a's of aspartate and glutamate residues with an error of 0.79.[53] A benchmark study on a large pK_a dataset of 100 proteins consisting of over 700 titratable residues showed that PROPKA has an accuracy of 85% within 1 pK unit for surface residues;[55] however, this dataset did not include the pK_a of any cysteine residues. PROPKA 3.1[54] was used for the pK_a's reported here.

2.3.3 Explicit Solvent Methods

For the models with an explicit representation of the solvent, cysteine pK_a's were determined by calculating the shift in pK_a from a reference value (i.e., ΔpK_a),

$$pK_a(\text{residue}) = pK_a(\text{reference}) + \Delta pK_a\,(\text{solvent} \rightarrow \text{protein}) \tag{2.1}$$

Here, pK_a(reference) refers to the reference pK_a value of the residue in solution. ΔpK_a results from the difference in the intermolecular interactions experienced by the protonated and deprotonated states of the amino acid side chain in the protein relative to the interaction they experience in bulk solution.

The cysteine pK_a shifts are calculated using the thermodynamic scheme illustrated in Figure 2.3, introduced by Warshel and coworkers.[86] The free energy difference between the protonated and deprotonated states of a residue are calculated in the protein environment and in a reference model system. The reference system is typically a short, blocked peptide that contains the ionizable residue of interest. An accurate experimental pK_a of this reference value, pK_a(reference), must be known so it can be used to calculate the absolute pK_a's in the protein using eq 2.1.

The relative Gibbs energies of the protonated and deprotonated states for both the residue in the protein and in the reference system are computed using free energy calculation methods. The difference between these two values gives the Gibbs energy of deprotonation of the protein residue relative to the reference value ($\Delta\Delta$G), eq 2.2.

$$\Delta\Delta G = \Delta G\,(protein) - \Delta G\,(reference) \tag{2.2}$$

In turn, the relative pK_a shift can be calculated from the relative Gibbs energy,

$$\Delta pK_a = \frac{\Delta\Delta G}{2.303RT} \tag{2.3}$$

where R and T are the gas constant and simulation temperature, respectively. In this study, the relative Gibbs energies are calculated using replica-exchange thermodynamic integration simulations of the proteins in an explicit solvent.

The reference system used in this study was a blocked pentapeptide with the sequence: ACE-Ala-Ala-Cys-Ala-Ala-NH_2. The experimental pK_a of the cysteine residue in this peptide is 8.55 ± 0.03.[84] Acetyl (ACE) and amine (NH_2) functionalities were used as capping groups for the N and C termini, respectively, to avoid artifacts from charged termini.

2.3.4 Thermodynamic Integration

Thermodynamic integration (TI)[87] provides a means of calculating the relative free energies of the protonated and deprotonated states of the protein. The free energy difference of the two states ($\Delta G_{A\to B}$) is expressed as an integral of the derivative of potential energy as a function of a continuously varying parameter, λ. The formal expression for the free energy difference between the two given states, labeled here as A and B, of a system is,

$$\Delta G_{A\to B} = \int_{\lambda=0}^{\lambda=1} \left(\frac{\partial G}{\partial \lambda}\right) \mathrm{d}\lambda = \int_{\lambda=0}^{\lambda=1} \left\langle \frac{\partial \mathcal{V}(\lambda)}{\partial \lambda} \right\rangle_{\lambda} \mathrm{d}\lambda \tag{2.4}$$

where λ is the reaction coordinate connecting states A and B. $\lambda = 0$ corresponds

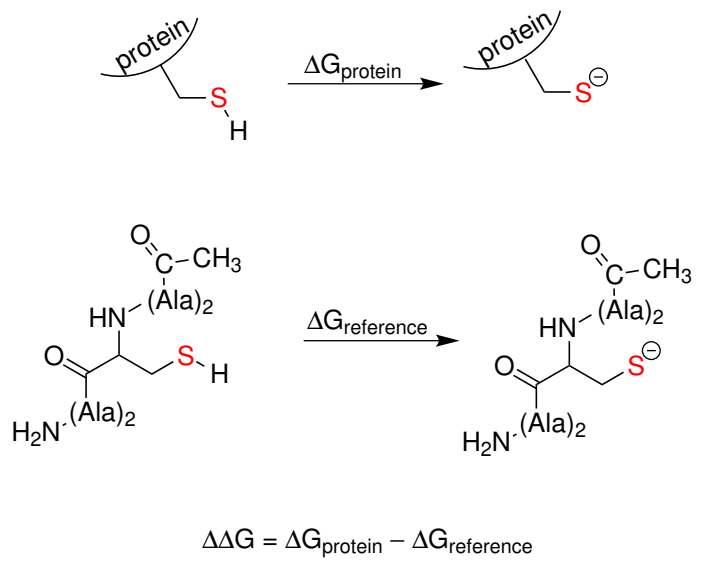

Figure 2.3: Alchemical transformation used in thermodynamic integration calculations of cysteine pK_a shifts. Top — Deprotonation reaction of cysteine residue in protein environment; Bottom — Deprotonation reaction of cysteine residue in reference model pentapeptide; $\Delta\Delta G$ refers to the relative free energy difference between the two reaction schemes.

to the reference or initial state (A), while $\lambda = 1$ corresponds to the final state (B). λ values in-between 0 and 1 represent an alchemical hybrid of the two states. The term $\mathcal{V}(\lambda)$ refers to the potential energy function along the reaction coordinate and $\langle\ \rangle_\lambda$ represents an ensemble average at a particular λ value. For the purposes of pK_a calculations, TI is used to calculate the free energy difference between the protonated and deprotonated forms of the amino acid in the protein.

A limitation of TI is that long simulation times can be required for the calculated free energies to converge. For instance, the anionic thiolate state inside an active site can have a range of hydration structures, but transitions between these states can be rare within the nanosecond time scales of conventional TI simulations. The infrequency of these transitions causes the sampling of these windows to converge slowly.

Replica-exchange methods[88–90] can improve the rate of convergence for TI simulations, allowing the relative free energies to be calculated with less sampling error. With these methods, each simulation (a.k.a., replica) with a value of λ is run in parallel. Periodically, exchanges are attempted between neighboring replicas. The

acceptance probability of the exchanges is,[91]

$$P_{acc} = \min\left[1, \exp\left(\frac{-\{[\mathcal{H}_{\lambda_i}(\mathbf{r}_j, \mathbf{p}_j) + \mathcal{H}_{\lambda_j}(\mathbf{r}_i, \mathbf{p}_i)] - [\mathcal{H}_{\lambda_i}(\mathbf{r}_i, \mathbf{p}_i) + \mathcal{H}_{\lambda_j}(\mathbf{r}_j, \mathbf{p}_j)]\}}{k_B T}\right)\right] \tag{2.5}$$

where $\mathcal{H}_{\lambda_i}$ is the Hamiltonian of the system for the i^{th} replica. $\mathbf{r}_i$ and $\mathbf{p}_i$ are the coordinates and momenta of the particles of the i^{th} replica, respectively.

These exchanges improve the efficiency of the configurational space sampling of the windows by allowing different configurations to be accessed by exchanges rather than dynamical transitions. Meng et al. showed that this type of Hamiltonian replica exchange dramatically improved the convergence of a free energy perturbation calculation of the pK_a of Asp26 in thioredoxin.[92]

For the calculation of cysteine pK_a's with an explicit solvent model, we used the all-atom Replica-Exchange Thermodynamic Integration (RETI) technique to calculate the relative Gibbs energy of the thiol and thiolate states (Fig. 2.4). The neutral cysteine was defined as the initial state ($\lambda = 0.0$). The final state ($\lambda = 1.0$) corresponds to the negatively charged thiolate anion. To maintain a neutral simulation cell, a chloride ion restrained in solution is neutralized in the final state of the system. The only difference in the potential energy function, $\mathcal{V}$, of the two states are the charges, and Lennard-Jones parameters are linearly interpolated between the initial and final states as $\lambda = 0 \rightarrow 1$,

$$\mathcal{V}(\lambda) = (1 - \lambda)\mathcal{V}_{\lambda=0} + \lambda\mathcal{V}_{\lambda=1}. \tag{2.6}$$

Sample topology files showing the charges and Lennard-Jones parameters for the cysteine thiol/thiolate states are provided (see Appendix A).

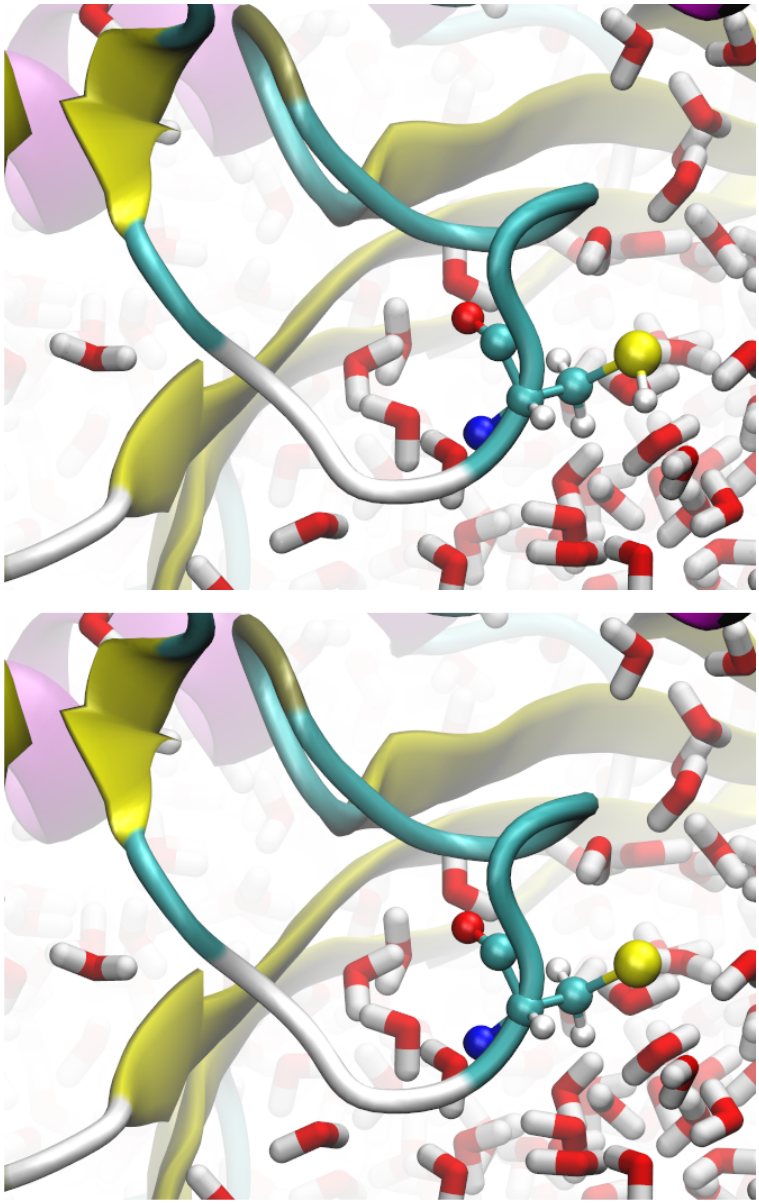

Figure 2.4: Representative configuration of Cys283 from an all-atom RETI simulation of human muscle creatine kinase (PDB ID: `1I0E`). Top – Explicit representation of active site thiol cysteine ($\lambda = 0$); Bottom – Explicit representation of active site thiolate cysteine ($\lambda = 1$).

2.3.5 Technical Details of RETI Calculations

Models for the all-atom simulations of proteins in our test set were constructed from structures deposited in the protein data bank. These structures were used to generate models using the CHARMM36[61] and AMBER ff99SB-ILDNP[62] force fields. The editconf module within the GROMACS software package was used to define a periodic cubic box with the protein centred and placed at least 1.0 nm from the edge of the box. Solvation of the protein within the periodic cubic box was accomplished using the solvate module within the GROMACS software package. Na^+ and Cl^- ions were added to the simulation cell such that the unit cell had no net charge and the ion concentration was approximately 0.10 M. Water molecules were represented using the TIP3P model.[93]

The simulation temperature and pressure were kept constant at 298.15 K and 100 kPa, respectively, by the velocity-rescaling thermostat[94] and the Parrinello–Rahman barostat.[95,96] Covalent bonds to hydrogen were constrained with the LINear Constraint Solver (LINCS) algorithm.[97] Long range electrostatics were treated by the Particle Mesh Ewald (PME) method.[98] A grid spacing of 1.0 Å was used for all simulation cells and a cut-off distance of 1.0 nm was chosen for both the real space Coulombic and Lennard-Jones interactions.

The initial structure was subjected to a steepest descent energy minimization to eliminate steric atomic clashes or structural irregularities that may exist within the constructed model system. Afterwards, 10–20 ns equilibration simulations were performed using a simulation in the canonical ensemble (NVT) followed by a simulation in the isothermal-isobaric (NpT) ensemble. These simulations used a velocity-rescaling thermostat[94] with a reference temperature 298.15 K and the Parrinello–Rahman barostat[95,96] with a reference pressure of 100 kPa. The time step was 2 fs.

After equilibration, the relative free energies of the thiol and thiolate states were calculated using RETI. Each simulation was comprised of 11 windows with λ= 0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 1.0. Exchanges were attempted between neighboring replicas every 1 ps. The average exchange probability between replicas was in the 0.1–0.2 range. The RETI simulations were 12 ns in length, with the first 2 ns discarded for equilibration. All free energy calculations and molecular dynamics (MD) simulations were performed using the GROMACS 4.5.4 software package.[99] Gibbs energies were calculated from the RETI data using g_wham.[100]

2.4 Results and Discussion

Table 2.2: Comparison of calculated cysteine pK_a's using the implicit and explicit solvation methods for the residues in the test set.

Abbrev.	Exptl. pK_a	H++	MCCE	PROPKA	CHARMM	AMBER
α-1-AT	6.86 ± 0.05	7.28	8.29	9.06	7.64 ± 0.35	9.35 ± 0.67
AhpC	5.94 ± 0.10	9.38	9.09	9.14	8.19 ± 0.46	9.66 ± 0.61
ACBP-M46C	8.20 ± 0.10	8.77	8.80	9.03	8.69 ± 0.16	6.96 ± 0.34
ACBP-S65C	9.00 ± 0.10	8.81	9.39	9.57	8.13 ± 0.20	7.55 ± 0.56
ACBP-T17C	9.80 ± 0.10	8.39	8.76	8.87	8.97 ± 0.44	10.34 ± 0.85
DJ-1[1]	5.40 ± 0.10	11.25	12.55	12.28	5.40 ± 0.98	8.80 ± 0.34
HMCK	5.60 ± 0.10	9.10	6.82	10.41	7.06 ± 0.38	7.09 ± 1.15
HMCK-S285A	6.70 ± 0.10	9.31	6.60	11.21	6.46 ± 0.79	9.38 ± 0.39
Mb-G124C	6.53 ± 0.05	8.06	8.46	8.35	7.95 ± 0.77	8.84 ± 0.60
Mb-A125C	8.43 ± 0.03	8.25	8.79	9.15	8.65 ± 0.77	8.45 ± 0.47
MmsrA[2]	7.20 ± 0.20	>12	16.29	13.09	7.45 ± 0.63	10.68 ± 0.99
MmsrA-E115Q[3]	8.20 ± 0.10	>12	15.42	11.36	7.01 ± 0.66	13.24 ± 0.68
O^6-AGT	5.30 ± 0.20	9.51	8.34	10.57	8.67 ± 0.52	10.90 ± 0.67
papain[4]	3.32 ± 0.01	9.32	8.84	10.50	4.44 ± 0.57	4.95 ± 0.83
ppΩ[5]	2.88 ± 0.02	9.44	7.57	7.45	0.71 ± 0.83	-0.16 ± 0.73
PTP1B	5.57 ± 0.12	3.93	-0.66	8.50	1.18 ± 0.40	1.31 ± 0.27
YopH	4.67 ± 0.15	3.97	-0.98	7.55	2.89 ± 0.71	4.63 ± 0.83
YopH-H402A	7.35 ± 0.04	3.29	-0.63	7.54	0.25 ± 1.03	1.39 ± 0.39

The protonation states of all residues in the pK_a test set were determined automatically for the methods using implicit solvent approach (i.e., H++, MCCE, PROPKA). CHARMM and AMBER simulations on the other hand, were performed with select residues in non-standard protonation states, in accordance with experimental pK protonation state conditions.

[1] DJ-1: Glu18 protonated;
[2] MmsrA: Glu115, Asp150, His206 protonated;
[3] MmsrA-E115Q: Asp150, His206 protonated;
[4] Papain: His159 protonated;
[5] ppΩ: His159 protonated.

2.4.1 Implicit Solvent Methods

Table 2.3 lists the RMSDs of the predicted pK_a's from experiment using the different pK_a prediction methods. The implicit solvent methods were generally inaccurate for predicting cysteine pK_a's (Figure 2.5). For each of these methods, the RMSD for cysteine pK_a prediction was greater than 3 pK units. Among the three implicit solvent methods used, H++ reported cysteine pK_a values with the smallest deviation from experiment (RMSD = 3.41). The pK_a's calculated using MCCE had the largest deviations from the experimental values (RMSD = 4.72). We note that because the

size of our test set is small, the uncertainty for the RMSDs is large,[101] so the relative accuracy of these methods cannot be definitively assigned from these calculations.

PROPKA has a modest RMSD, but this is largely because all pK_a's are predicted within 7.4–13.1 range, so the margin of error is smaller than for some of the other methods that predict more extreme pK_a values in some cases. The accuracy of PROPKA for cysteine pK_a prediction is generally poor; the pK_a of residues are overestimated in 16 of the 18 cases and this method is generally unreliable for predicting the direction or magnitude of the pK_a shift. Even cysteine proteases that have $Cys^{(-)}$—$His^{(+)}$ ion pair are predicted to have pK_a's in the 7–10 range characteristic of non-catalytic cysteines, indicating that the effect of thiolate stabilization by charge–charge interactions within the protein is underestimated.

The pK_a's of papain, ppΩ, MmsrA, and DJ-1 are significantly overestimated by all the implicit solvent methods. The protic cysteine residue in these proteins are in close contact with another acidic residue, such as Glu or His. In order for these methods to predict the pK_a of the cysteine residue correctly, it must correctly predict the pK_a of these coupled residues and the effect of this residue on the cysteine. This type of electrostatic coupling with residues that are not in their standard protonation states can be a challenge for these methods.

Table 2.3: RMSDs (σ) and associated error interval reported at 80% confidence limit for cysteine pK_a methods. The error bounds on the RMSDs were calculated using eq. (74) of Ref. 101. H++ does not report absolute pK_a's for Cys-72 in MMsrA & MmsrA-E115Q, so these values are not included in the RMSD (n=18) calculation.

Method	**RMSD (n=18)**		**RMSD (n=16)**	
	σ	Range for $\chi^2_{80\%}$	σ	Range for $\chi^2_{80\%}$
H++	–	–	3.41	$2.72 < \sigma < 4.33$
MCCE	4.72	$3.82 < \sigma < 5.90$	4.08	$3.26 < \sigma < 5.18$
PROPKA	3.90	$3.15 < \sigma < 4.88$	3.78	$3.02 < \sigma < 4.80$
CHARMM	2.40	$1.94 < \sigma < 3.00$	2.52	$2.01 < \sigma < 3.20$
AMBER	3.20	$2.59 < \sigma < 4.00$	3.03	$2.42 < \sigma < 3.85$
Null-model	2.74	$2.22 < \sigma < 3.43$	2.88	$2.30 < \sigma < 3.66$

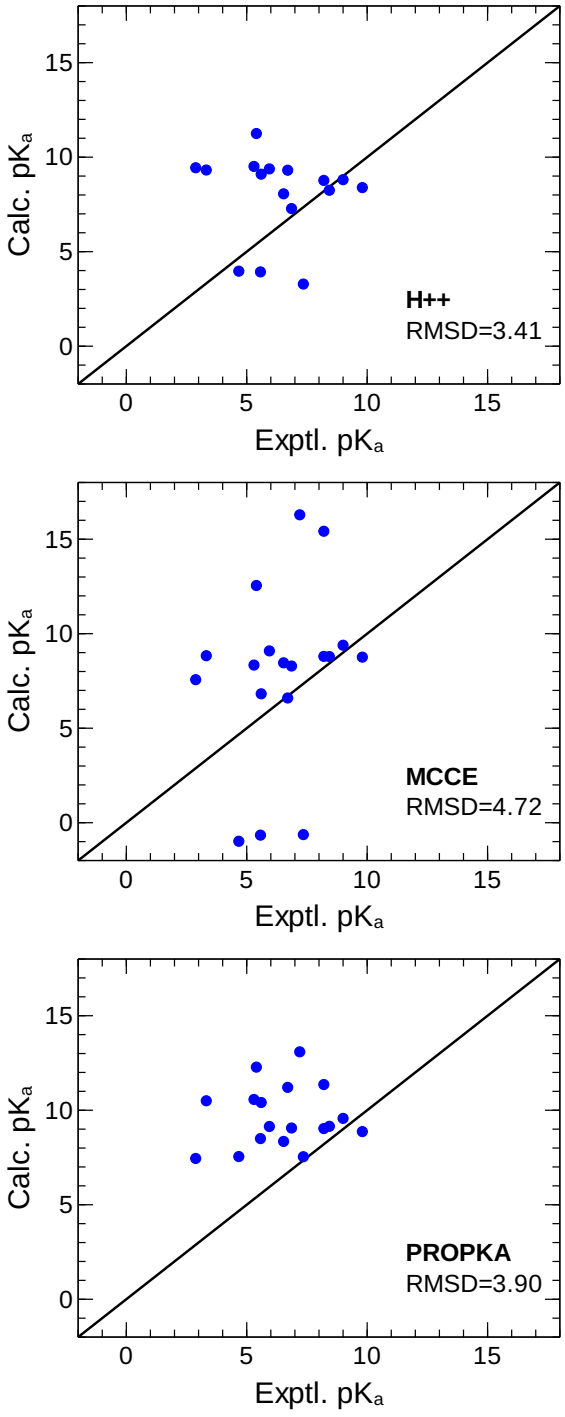

Figure 2.5: Correlation between experimental and calculated cysteine pK_a's using the H++, MCCE, and PROPKA implicit solvent pK_a prediction methods.

MCCE and H++ methods are able to correctly predict that some catalytic cysteine residues will have a depressed pK_a, but the magnitude of this effect tends to be overestimated, especially by the MCCE program. This is certainly the case for the predicted pK_a's of YopH and PTP1B, which are predicted to be significantly lower than reported experimental pK_a values. The catalytic cysteines in these proteins neighbor charged residues, which can stabilize the thiolate state of the cysteine residue. The underestimation of the pK_a's of these proteins by these methods suggests that the pK_a-lowering effect by nearby cationic residues is not fully accounted for. In general, these methods appear to underestimate the stability of cysteine groups with coupled protonation states to other ionizable residues (e.g., His and Arg), leading to significant disparities between predicted and experimentally reported cysteine pK_a values.

Mouse methionine sulfoxide reductase A (MmsrA) is a particularly challenging case for these methods; the experimental pK_a is 7.4 but MCCE and PROPKA predict values of 16.3 and 13.1, respectively. H++ does not report values of cysteine pK_a's that it deems to be greater than 12, so it only reports that these pK_a's are > 12. The catalytic cysteine residue (Cys72) in MmsrA interacts with a neighboring protonated Glu-115 residue.[77] These implicit solvent methods predict this residue to be anionic, so the pK_a is predicted to be shifted higher rather than lower. If MmsrA and its E115Q mutant are excluded ($n = 16$), the RMSDs of predicted pK_a's using the MCCE and PROPKA methods are reduced to 4.08 and 3.78, respectively.

2.4.2 Explicit Solvent Methods

The explicit solvent RETI pK_a calculations show a significant improvement over the implicit solvent methods (Table 2.3). The accuracy of these calculations are sensitive to the force field used; the RMSD of the RETI calculations with the CHARMM36 force field are 2.40 while the RMSD for the calculations with the AMBER force field are 3.20 (Table 2.3). Figure 2.6 shows the correlation between the calculated and experimental pK_a's using these methods. The most significant improvement of these methods over the implicit solvent methods are in predicting depressed pK_a's. It should be noted that these RMSDs are calculated from a sample size ($n = 18$). With an 80% confidence limit, the level of uncertainty within the calculated RMSDs is too large to conclusively say that these simulations are more accurate than the null-model; but we can conclude that the CHARMM36 calculations and the null-model outperform the

implicit solvent methods.

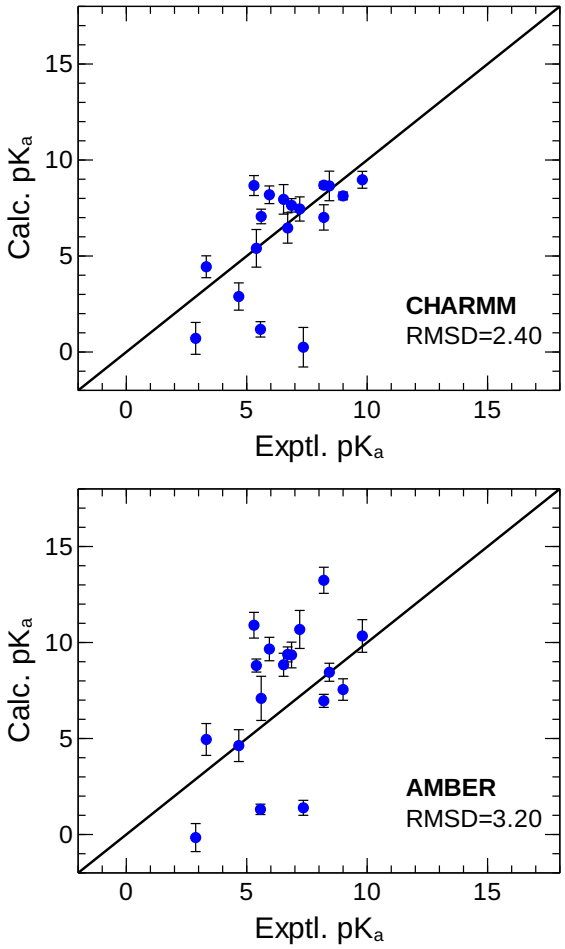

Figure 2.6: Correlation between experimental and calculated cysteine pK$_a$'s using the explicit solvent RETI method with the CHARMM36 and AMBER ff99SB-ILDNP force fields.

The calculation of the pK$_a$ of Cys72 of MmsrA was greatly in error when the other residues of the protein were assigned standard protonation states. The RETI simulations with the CHARMM and AMBER force fields predicted pK$_a$'s of 11.9 and 10.7, respectively, while the experimental value is 7.2. The predictions with the CHARMM force field were improved considerably when nearby residues, Asp150 and His206,

were also protonated; with these modifications, the predicted pK_a is decreased to 7.4, in good agreement with experiment. Experiments or more sophisticated simulations would be needed to rigorously show that these assignments are correct.

These methods perform surprisingly poorly on AhpC. The reported pK_a of Cys46 in AhpC is 5.94.[69] The thiolate state of this residue is stabilized by the nearby cationic charge of Arg119, although the RETI-predicted pK_a's are 8.19 and 9.66 for the CHARMM and AMBER force fields, respectively. The effect of this arginine–thiolate salt bridge appears to be underestimated, or the effect of the more distant destabilizing interaction with Glu49 has too large of an impact. Another possibility is that there is a significant structural change in this protein between the protonated and deprotonated states that is not captured in these simulations. The protein backbone RMSD of the final coordinates of the $\lambda = 1$ window is only 4.27 Å, indicating that the thiolate state did not undergo a large conformational change on the timescale of our simulation (see Table A.3 in Appendix A).

These methods also perform poorly for O^6-AGT; the pK_a should be depressed to 5.30, but both the CHARMM and AMBER models predict elevated pK_a's (8.67 and 10.90, respectively). If the simulations are performed with the nearby His146 residue in its cationic state, the pK_a's are underestimated, with values of 1.44 and 4.04, respectively. This suggests that the His protonation state affects the pK_a results and that His146 is actually only protonated a fraction of the time, so the actual pK_a is an intermediate between these two extremes. The protonation states of other ionizable amino acids in the protein remains constant in this type of RETI calculation, which limits the accuracy of the calculations in cases like this. Non-standard protonation states of ionizable residues used in the RETI simulations are presented in Table 2.4.

Table 2.4: Cysteine test set proteins with residues of non-standard protonation states for RETI simulations.

Abbrev.	**Cys Residue**	**Exptl. pK_a**	**Protonated Residue**
DJ-1	106	5.40 ± 0.10	Glu18
MmsrA	72	7.20 ± 0.20	Glu115, Asp150, His206
MmsrA-E115Q	72	8.20 ± 0.01	Asp150, His206
papain	25	3.32 ± 0.01	His159
ppΩ	25	2.88 ± 0.02	His159

2.4.3 Opportunities for Improvement

The limited accuracy of all these methods suggest there is considerable room for improvement for cysteine pK_a prediction methods. The poorer performance of these methods for cysteine residues compared to surface glutamic and aspartic acid residues may reflect a greater complexity in the acid-base chemistry of cysteine due to effects like induced polarization and coupled protonation states. It may be possible to address these issues using more realistic models, more extensive simulations, and validation on a more extensive set of cysteine pK_a's.

The three implicit solvent methods are significantly less accurate for cysteine residues than for Glu and Asp residues. For example, the RMSD of pK_a's estimated by PROPKA for a test set consisting of 201 Glu and Asp residues was only 0.79.[53] These methods are partially empirical and were primarily developed for Asp and Glu amino acids. It is possible that the performance of these methods for cysteine residues could be improved if they are redeveloped specifically for cysteine residue pK_a's. The pK_a's of a small number of cysteine residues calculated using PROPKA 2.0 were predicted reasonably well,[33] so it may be possible to improve current versions of PROPKA by adjusting parameters like distance cutoffs. The partial or complete neglect of protein dynamics is another serious limitation for these methods.

The explicit solvent models are systematically more accurate, so these are a more promising base for further development. The disparity between the AMBER and CHARMM results indicates that the calculated pK_a's are sensitive to the force field used in the simulations. The conformational ensemble sampled from the simulation can depend strongly on the force field used.[102–104] The thiol/thiolate parameters will also have a large effect on the computed pK_a. In particular, the thiolate parameters in these force fields have not been optimized for pK_a calculations. It is possible that reparameterization of these force fields could yield more accurate results, although there was no systematic trend in the deviations from experiment, so there is no obvious parameter to adjust.

Polarizable force fields are one possible means to improve the accuracy of these calculations. Sulfur is highly polarizable[105–107] and a more realistic description of the electrostatic environment of the protein could provide a significant improvement.[108] Perhaps more significantly, the thiolate is a diffuse, polarizable anion,[109] so induced

polarization could have a large effect on the stability of the deprotonated state. Alternatively, QM/MM methods have shown promise in the calculation of pK_a's, including cysteine.[31,34,35] Ab initio methods like SCC-DFTB could be particularly useful for cysteine pK_a's because the experimental data about thiolate structure and energetics is relatively limited, which makes it difficult to parameterize an empirical force field.

The simulations performed using the thermodynamic integration technique were limited to 12 ns simulations, which is generally sufficient to sample the relative energies if the conformation of the protein is generally preserved. A change in the protonation state of a residue can result in significant changes to the conformation of a protein.[110,111] Capturing such a large conformational change in a TI simulation would require very long simulations or the use of more sophisticated enhanced sampling methods. The RMSDs of the protein backbone of the RETI simulations ranged from 1–4 Å for both the $\lambda = 0$ and $\lambda = 1$ windows (see Table A.3 in Appendix A). This suggests that our simulation times were reasonable for sampling the relative Gibbs energies of the two states given that no major rearrangements occurred, but also that if a large rearrangement should have occurred, it did not occur within the timescale of our simulations.

Another challenge is that there are relatively few experimentally-determined cysteine pK_a's in comparison to the pK_a's of acidic amino acids, like glutamic and aspartic acid. This contributed to a large uncertainty in the RMSDs relative to experiment for these methods. Furthermore, the experimentally-reported cysteine pK_a's used in the test set were determined through a wide range of physical techniques, such as relative reaction kinetics, spectroscopic titration, and microcalorimetry, so the test set may include inconsistent values.

Lastly, the protonation state of cysteine side chains are often coupled to the protonation states of other residues. In particular, histidine has a similar intrinsic pK_a to cysteine ($pK_a \approx 6$) and the protonation states of catalytic cysteines are often lowered by the electrostatic interaction with a nearby cationic histidine residue. A typical example of this activity occurs in the enzyme papain, where the active site cysteine (Cys25) thiolate anion is stabilized by a nearby imidazolium cationic histidine (His159) forming a zwitterionic pair (Fig. 2.7).[112]. In the RETI simulations presented here, the protonation states of all residues other than the target cysteine are fixed. In other words, the charges of other ionizable residues are constant throughout

these simulations. As such, contributions of other protonation states are not captured by these simulations. These effects are particularly significant for catalytic cysteines because their pK_a's are often coupled to other residues.

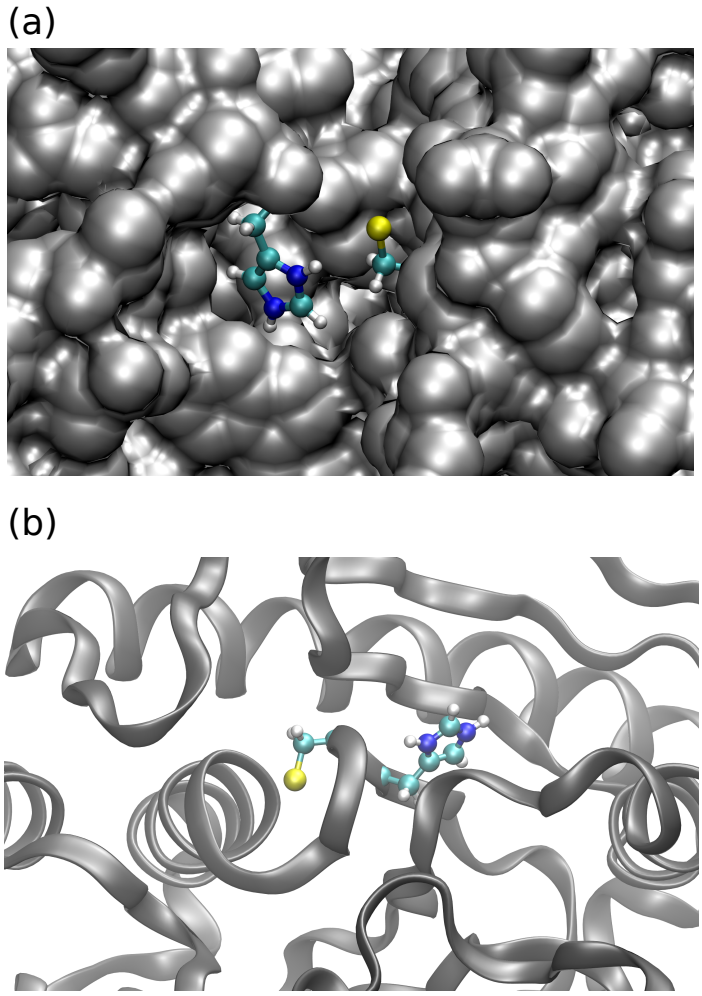

Figure 2.7: Structures of (a) Cys25–His159 ion pair based on the protein crystal structure (PDB ID: 1PPN). Salt bridge between the protonated histidine side chain and the deprotonated cysteine side chain results in a large reduction of the pK_a of Cys25 (pK_a(exptl)=3.32).[16] and (b) the pK_a-coupled residues Cys403 and His402 based on the protein crystal structure (PDB ID: 1YTW). The protein structure precludes a Cys–His salt bridge, but a through-space Coulombic interaction between the deprotonated cysteine and the protonated histidine effects a large decrease in the cysteine pK_a (WT pK_a(exptl)=4.67 vs. H402A mutant pK_a=7.35).[17]

Newly-developed constant-pH methods offer promising solutions to the limitations of pK_a prediction methods based on dual-state free energy calculations. A method by Mongan et al. allows a protein described by the Generalized-Born model to transition between protonation states via Monte Carlo moves.[113] The pK_a of a residue can be calculated from the calculated "titration" curve. These methods can be combined with temperature-exchange molecular dynamics to improve configurational sampling.[92] A hybrid explicit solvent/implicit solvent scheme has also been developed where the MD simulations are performed with an explicit solvent but the transition probabilities between protonation states are calculated using an implicit solvent model.[114] In a similar vein, Chen and Roux have developed a new method that attempts transitions between protonation states by non-equilibrium Monte Carlo moves, which allows an explicit solvent model to be used consistently.[115] Constant-pH MD methods, where the protonation state varies as a dynamical variable, also show promise.[116–119] The maturation and widespread implementation of these methods could be particularly important for the quantitative calculation of cysteine pK_a's.

2.5 Conclusions

A test set of 18 cysteine side chain pK_a's in 12 proteins was selected to evaluate how effective computational methods are for predicting the pK_a's of these residues. Three implicit solvent methods were compared: H++, MCCE, and PROPKA. An explicit solvent method that uses replica-exchange thermodynamic integration was also tested with the CHARMM36 and AMBER ff99SB-ILDNP force fields.

The implicit solvent methods were found to be highly unreliable, with RMSDs > 3 pK units. These methods are less accurate for cysteine residues than for other types of ionizable side chains. H++ performed incrementally better than the other implicit methods, with an RMSD of 3.41.

The explicit solvent methods were found to be significantly more accurate. The results from the CHARMM36 calculations showed the highest accuracy. Nevertheless, the accuracy of the simulations using the CHARMM36 force field was limited, with an RMSD of 2.40. This is considerably less accurate than pK_a predictions for other

types of side chains. Improvements in the models for thiolate state of the cysteine side chain and descriptions of the protonation states of pK_a-coupled residues are areas where there are significant opportunities to improve the accuracy of cysteine pK_a predictions.

Bibliography

[1] Zeida, A.; Guardia, C. M.; Lichtig, P.; Perissinotti, L. L.; Defelipe, L. A.; Turjanski, A.; Radi, R.; Trujillo, M.; Estrin, D. A. Thiol Redox Biochemistry: Insights from Computer Simulations. *Biophys. Rev.* **2014**, *6*, 27–46.

[2] Marino, S. M.; Gladyshev, V. N. Cysteine Function Governs Its Conservation and Degeneration and Restricts Its Utilization on Protein Surfaces. *J. Mol. Biol.* **2010**, *404*, 902–916.

[3] Jensen, K. S.; Hansen, R. E.; Winther, J. R. Kinetic and Thermodynamic Aspects of Cellular Thiol–Disulfide Redox Regulation. *Antioxid. Redox Signaling* **2009**, *11*, 1047–1058.

[4] Rulıšek, L.; Vondrášek, J. Coordination Geometries of Selected Transition Metal Ions (Co^{2+}, Ni^{2+}, Cu^{2+}, Zn^{2+}, Cd^{2+}, and Hg^{2+}) in Metalloproteins. *J. Inorg. Biochem.* **1998**, *71*, 115–127.

[5] Chapman, H. A.; Riese, R. J.; Shi, G.-P. Emerging Roles for Cysteine Proteases in Human Biology. *Annu. Rev. Physiol.* **1997**, *59*, 63–88.

[6] Giles, N. M.; Giles, G. I.; Jacob, C. Multiple Roles of Cysteine in Biocatalysis. *Biochem. Biophys. Res. Commun.* **2003**, *300*, 1–4.

[7] Kim, H.-J.; Ha, S.; Lee, H. Y.; Lee, K.-J. ROSics: Chemistry and Proteomics of Cysteine Modifications in Redox Biology. *Mass Spectrom. Rev.* **2015**, *34*, 184–208.

[8] Roos, G.; Foloppe, N.; Messens, J. Understanding the pKa of Redox Cysteines: The Key Role of Hydrogen Bonding. *Antioxid. Redox Signaling* **2013**, *18*, 94–127.

[9] Otto, H.-H.; Schirmeister, T. Cysteine Proteases and Their Inhibitors. *Chem. Rev.* **1997**, *97*, 133–172.

[10] Denu, J. M.; Dixon, J. E. Protein Tyrosine Phosphatases: Mechanisms of Catalysis and Regulation. *Curr. Opin. Chem. Biol.* **1998**, *2*, 633–641.

[11] Kolmodin, K.; Åqvist, J. The Catalytic Mechanism of Protein Tyrosine Phosphatases Revisited. *FEBS Letters* **2001**, *498*, 208–213.

[12] Carmi, C.; Mor, M.; Petronini, P. G.; Alfieri, R. R. Clinical Perspectives for Irreversible Tyrosine Kinase Inhibitors in Cancer. *Biochem. Pharmacol.* **2012**, *84*, 1388–1399.

[13] Schwartz, P. A.; Kuzmic, P.; Solowiej, J.; Bergqvist, S.; Bolanos, B.; Almaden, C.; Nagata, A.; Ryan, K.; Feng, J.; Dalvie, D.; Kath, J. C.; Xu, M.; Wani, R.; Murray, B. W. Covalent EGFR Inhibitor Analysis Reveals Importance of Reversible Interactions to Potency and Mechanisms of Drug Resistance. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 173–178.

[14] Zhang, J.; Yang, P. L.; Gray, N. S. Targeting Cancer with Small Molecule Kinase Inhibitors. *Nature Rev. Cancer* **2009**, *9*, 28–39.

[15] Smith, J. M.; Rowley, C. N. Automated computational screening of the thiol reactivity of substituted alkenes. *J. Comput.-Aided Mol. Des.* **2015**, *29*, 725–735.

[16] Pinitglang, S.; Watts, A. B.; Patel, M.; Reid, J. D.; Noble, M. A.; Gul, S.; Bokth, A.; Naeem, A.; Patel, H.; Thomas, E. W.; Sreedharan, S. K.; Verma, C.; Brocklehurst, K. A Classical Enzyme Active Center Motif Lacks Catalytic Competence Until Modulated Electrostatically. *Biochemistry* **1997**, *36*, 9968–9982.

[17] Zhang, Z. Y.; Dixon, J. E. Active Site Labeling of the Yersinia Protein Tyrosine Phosphatase: The Determination of the pKa of the Active Site Cysteine and the Function of the Conserved Histidine 402. *Biochemistry* **1993**, *32*, 9340–9345.

[18] Lohse, D. L.; Denu, J. M.; Santoro, N.; Dixon, J. E. Roles of Aspartic Acid-181 and Serine-222 in Intermediate Formation and Hydrolysis of the Mammalian Protein-Tyrosine-Phosphatase PTP1. *Biochemistry* **1997**, *36*, 4568–4575.

[19] Bulaj, G.; Kortemme, T.; Goldenberg, D. P. Ionization–Reactivity Relationships for Cysteine Thiols in Polypeptides. *Biochemistry* **1998**, *37*, 8965–8972.

[20] Salsbury, F. R.; Knutson, S. T.; Poole, L. B.; Fetrow, J. S. Functional Site Profiling and Electrostatic Analysis of Cysteines Modifiable to Cysteine Sulfenic Acid. *Protein Sci.* **2008**, *17*, 299–312.

[21] Nakamura, H. Roles of Electrostatic Interaction in Proteins. *Q. Rev. Biophys.* **1996**, *29*, 1–90.

[22] Perutz, M. F. Electrostatic Effects in Proteins. *Science* **1978**, *201*, 1187–1191.

[23] Warshel, A. Electrostatic Basis of Structure-Function Correlation in Proteins. *Acc. Chem. Res.* **1981**, *14*, 284–290.

[24] Grauschopf, U.; Winther, J. R.; Korber, P.; Zander, T.; Dallinger, P.; Bardwell, J. C. Why is DsbA Such an Oxidizing Disulfide Catalyst? *Cell* **1995**, *83*, 947–955.

[25] Dyson, H. J.; Jeng, M.-F.; Tennant, L. L.; Slaby, I.; Lindell, M.; Cui, D.-S.; Kuprin, S.; Holmgren, A. Effects of Buried Charged Groups on Cysteine Thiol Ionization and Reactivity in Escherichia coli Thioredoxin: Structural and Functional Characterization of Mutants of Asp 26 and Lys 57. *Biochemistry* **1997**, *36*, 2622–2636.

[26] Jacobi, A.; Huber-Wunderlich, M.; Hennecke, J.; Glockshuber, R. Elimination of All Charged Residues in the Vicinity of the Active-Site Helix of the Disulfide Oxidoreductase DsbA: Influence of Electrostatic Interactions on Stability and Redox Properties. *J. Biol. Chem.* **1997**, *272*, 21692–21699.

[27] Hansen, R. E.; Ostergaard, H.; Winther, J. R. Increasing the Reactivity of an Artificial Dithiol–Disulfide Pair through Modification of the Electrostatic Milieu. *Biochemistry* **2005**, *44*, 5899–5906.

[28] Thurlkill, R. L.; Grimsley, G. R.; Scholtz, J. M.; Pace, C. N. pK Values of the Ionizable Groups of Proteins. *Protein Sci.* **2006**, *15*, 1214–1218.

[29] Nelson, K. J.; Day, A. E.; Zeng, B.-B.; King, S. B.; Poole, L. B. Isotope-coded, Iodoacetamide-based Reagent to Determine Individual Cysteine pKa Values by Matrix-Assisted Laser Desorption/Ionization Time-of-flight Mass Spectrometry. *Anal. Biochem.* **2008**, *375*, 187–195.

[30] Combining Conformational Flexibility and Continuum Electrostatics for Calculating pKas in Proteins. *Biophys. J.* **2002**, *83*, 1731–1748.

[31] Li, G.; Cui, Q. pKa Calculations with QM/MM Free Energy Perturbations. *J. Phys. Chem. B* **2003**, *107*, 14521–14528.

[32] Kuhn, B.; Kollman, P. A.; Stahl, M. Prediction of pKa Shifts in Proteins Using a Combination of Molecular Mechanical and Continuum Solvent Calculations. *J. Comput. Chem.* **2004**, *25*, 1865–1872.

[33] Li, H.; Robertson, A. D.; Jensen, J. H. Very fast empirical prediction and rationalization of protein pKa values. *Proteins: Struct., Funct., Bioinf.* **2005**, *61*, 704–721.

[34] Riccardi, D.; Schaefer, P.; Cui, Q. pKa Calculations in Solution and Proteins with QM/MM Free Energy Perturbation Simulations: A Quantitative Test of QM/MM Protocols. *J. Phys. Chem. B* **2005**, *109*, 17715–17733.

[35] Jensen, J. H.; Li, H.; Robertson, A. D.; Molina, P. A. Prediction and Rationalization of Protein pKa Values Using QM and QM/MM Methods. *J. Phys. Chem. A* **2005**, *109*, 6634–6643.

[36] Gordon, J. C.; Myers, J. B.; Folta, T.; Shoja, V.; Heath, L. S.; Onufriev, A. H++: a server for estimating pKas and adding missing hydrogens to macromolecules. *Nucleic Acids Res.* **2005**, *33*, W368–W371.

[37] Spassov, V. Z.; Yan, L. A Fast and Accurate Computational Approach to Protein Ionization. *Protein Sci.* **2008**, *17*, 1955–1970.

[38] Huang, R.-B.; Du, Q.-S.; Wang, C.-H.; Liao, S.-M.; Chou, K.-C. A Fast and Accurate Method for Predicting pKa of Residues in Proteins. *Protein Eng. Des. Sel.* **2010**, *23*, 35–42.

[39] Tanford, C.; Kirkwood, J. G. Theory of Protein Titration Curves. I. General Equations for Impenetrable Spheres. *J. Am. Chem. Soc.* **1957**, *79*, 5333–5339.

[40] Gilson, M. K.; Honig, B. H. Calculation of Electrostatic Potentials in an Enzyme Active Site. *Nature* **1987**, *330*, 84–86.

[41] Bashford, D.; Karplus, M. pKa's of Ionizable Groups in Proteins: Atomic Detail From a Continuum Electrostatic Model. *Biochemistry* **1990**, *29*, 10219–10225.

[42] Tanford, C.; Roxby, R. Interpretation of Protein Titration Curves. Application to Lysozyme. *Biochemistry* **1972**, *11*, 2192–2198.

[43] Yang, A.-S.; Gunner, M. R.; Sampogna, R.; Sharp, K.; Honig, B. On the Calculation of pKas in Proteins. *Proteins: Struct., Funct., Bioinf.* **1993**, *15*, 252–265.

[44] Antosiewicz, J.; McCammon, J. A.; Gilson, M. K. Prediction of pH-dependent Properties of Proteins. *J. Mol. Biol.* **1994**, *238*, 415–436.

[45] Potter, M. J.; Gilson, M. K.; McCammon, J. A. Small Molecule pKa Prediction with Continuum Electrostatics Calculations. *J. Am. Chem. Soc.* **1994**, *116*, 10298–10299.

[46] Cramer, C. J.; Truhlar, D. G. General Parameterized SCF Model for Free Energies of Solvation in Aqueous Solution. *J. Am. Chem. Soc.* **1991**, *113*, 8305–8311.

[47] Jayaram, B.; Liu, Y.; Beveridge, D. L. A Modification of the Generalized Born Theory for Improved Estimates of Solvation Energies and pK Shifts. *J. Chem. Phys.* **1998**, *109*, 1465.

[48] Myers, J.; Grothaus, G.; Narayanan, S.; Onufriev, A. A Simple Clustering Algorithm Can Be Accurate Enough for Use in Calculations of pKs in Macromolecules. *Proteins: Struct. Funct. Bioinf.* **2006**, *63*, 928–938.

[49] Anandakrishnan, R.; Aguilar, B.; Onufriev, A. V. H++ 3.0: automating pK Prediction and the Preparation of Biomolecular Structures for Atomistic Molecular Modeling and Simulations. *Nucleic Acids Res.* **2012**, *40*, W537–W541.

[50] Alexov, E.; Gunner, M. Incorporating Protein Conformational Flexibility into the Calculation of pH-Dependent Protein Properties. *Biophys. J.* **1997**, *72*, 2075.

[51] Song, Y.; Mao, J.; Gunner, M. R. MCCE2: Improving protein pKa calculations with extensive side chain rotamer sampling. *J. Comput. Chem.* **2009**, *30*, 2231–2247.

[52] Bas, D. C.; Rogers, D. M.; Jensen, J. H. Very Fast Prediction and Rationalization of pKa Values for Protein–Ligand Complexes. *Proteins: Struct., Funct., Bioinf.* **2008**, *73*, 765–783.

[53] Olsson, M. H. M.; Søndergaard, C. R.; Rostkowski, M.; Jensen, J. H. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions. *J. Chem. Theory Comput.* **2011**, *7*, 525–537.

[54] Søndergaard, C. R.; Olsson, M. H. M.; Rostkowski, M.; Jensen, J. H. Improved Treatment of Ligands and Coupling Effects in Empirical Calculation and Rationalization of pKa Values. *J. Chem. Theory Comput.* **2011**, *7*, 2284–2295.

[55] Davies, M. N.; Toseland, C. P.; Moss, D. S.; Flower, D. R. Benchmarking pKa Prediction. *BMC Biochem.* **2006**, *7*, 18.

[56] Lee, A. C.; Crippen, G. M. Predicting pKa. *J. Chem. Inf. Model.* **2009**, *49*, 2013–2033.

[57] Simonson, T.; Carlsson, J.; Case, D. A. Proton Binding to Proteins: pKa Calculations with Explicit and Implicit Solvent Models. *J. Am. Chem. Soc.* **2004**, *126*, 4167–4180.

[58] Alexov, E.; Mehler, E. L.; Baker, N.; M. Baptista, A.; Huang, Y.; Milletti, F.; Erik Nielsen, J.; Farrell, D.; Carstensen, T.; Olsson, M. H. M.; Shen, J. K.; Warwicker, J.; Williams, S.; Word, J. M. Progress in the Prediction of pKa Values in Proteins. *Proteins: Struct., Funct., Bioinf.* **2011**, *79*, 3260–3275.

[59] Marino, S. M.; Gladyshev, V. N. Analysis and Functional Prediction of Reactive Cysteine Residues. *J. Biol. Chem.* **2012**, *287*, 4419–4425.

[60] Stanton, C. L.; Houk, K. N. Benchmarking pKa Prediction Methods for Residues in Proteins. *J. Chem. Theory Comput.* **2008**, *4*, 951–966.

[61] Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E. M.; Mittal, J.; Feig, M.; Alexander D. MacKerell, J. Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone ϕ, ψ and Side-Chain $\chi1$ and $\chi2$ Dihedral Angles. *J. Chem. Theory Comput.* **2012**, *8*, 3257–3273.

[62] Aliev, A. E.; Kulke, M.; Khaneja, H. S.; Chudasama, V.; Sheppard, T. D.; Lanigan, R. M. Motional Timescale Predictions by Molecular Dynamics Simulations: Case Study Using Proline and Hydroxyproline Sidechain Dynamics. *Proteins: Struct., Funct., Bioinf.* **2014**, *82*, 195–215.

[63] 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*, 235–242.

[64] Arnold, K.; Bordoli, L.; Kopp, J.; Schwede, T. The SWISS-MODEL Workspace: A Web-based Environment for Protein Structure Homology Modelling. *Bioinformatics* **2006**, *22*, 195–201.

[65] Elliott, P. R.; Pei, X. Y.; Dafforn, T. R.; Lomas, D. A. Topography of a 2.0 Å Structure of α1-antitrypsin Reveals Targets for Rational Drug Design to Prevent Conformational Disease. *Protein Sci.* **2000**, *9*, 1274–1281.

[66] Griffiths, S. W.; King, J.; Cooney, C. L. The Reactivity and Oxidation Pathway of Cysteine 232 in Recombinant Human α1-Antitrypsin. *J. Biol. Chem.* **2002**, *277*, 25486–25492.

[67] Jensen, K. S.; Pedersen, J. T.; Winther, J. R.; Teilum, K. The pKa Value and Accessibility of Cysteine Residues Are Key Determinants for Protein Substrate Discrimination by Glutaredoxin. *Biochemistry* **2014**, *53*, 2533–2540.

[68] Perkins, A.; Nelson, K. J.; Williams, J. R.; Parsonage, D.; Poole, L. B.; Karplus, P. A. The Sensitive Balance Between the Fully Folded and Locally Unfolded Conformations of a Model Peroxiredoxin. *Biochemistry* **2013**, *52*, 8708–8721.

[69] Nelson, K. J.; Parsonage, D.; Hall, A.; Karplus, P. A.; Poole, L. B. Cysteine pKa Values for the Bacterial Peroxiredoxin AhpC. *Biochemistry* **2008**, *47*, 12860–12868.

[70] Wilson, M. A.; Collins, J. L.; Hod, Y.; Ringe, D.; Petsko, G. A. The 1.1-Å Resolution Crystal Structure of DJ-1, the Protein Mutated in Autosomal Recessive Early Onset Parkinson's Disease. *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 9256–9261.

[71] Witt, A. C.; Lakshminarasimhan, M.; Remington, B. C.; Hasim, S.; Pozharski, E.; Wilson, M. A. Cysteine pKa Depression by a Protonated Glutamic Acid in Human DJ-1. *Biochemistry* **2008**, *47*, 7430–7440.

[72] Shen, Y.-Q.; Tang, L.; Zhou, H.-M.; Lin, Z.-J. Structure of Human Muscle Creatine Kinase. *Acta Crystallogr., Sect. D: Struct. Biol.* **2001**, *57*, 1196–1200.

[73] Wang, P.-F.; McLeish, M. J.; Kneen, M. M.; Lee, G.; Kenyon, G. L. An Unusually Low pKa for Cys282 in the Active Site of Human Muscle Creatine Kinase. *Biochemistry* **2001**, *40*, 11698–11705.

[74] Quillin, M. L.; Arduini, R. M.; Olson, J. S.; Phillips, G. N. High-Resolution Crystal Structures of Distal Histidine Mutants of Sperm Whale Myoglobin. *J. Mol. Biol.* **1993**, *234*, 140–155.

[75] Miranda, J. L. Position-Dependent Interactions Between Cysteine Residues and the Helix Dipole. *Protein Sci.* **2003**, *12*, 73–81.

[76] Lim, J. C.; Gruschus, J. M.; Ghesquière, B.; Kim, G.; Piszczek, G.; Tjandra, N.; Levine, R. L. Characterization and Solution Structure of Mouse Myristoylated Methionine sulfoxide reductase A. *J. Biol. Chem.* **2012**, *287*, 25589–25595.

[77] Lim, J. C.; Gruschus, J. M.; Kim, G.; Berlett, B. S.; Tjandra, N.; Levine, R. L. A Low pKa Cysteine at the Active Site of Mouse Methionine Sulfoxide Reductase A. *J. Biol. Chem.* **2012**, *287*, 25596–25601.

[78] Daniels, D. S.; Mol, C. D.; Arvai, A. S.; Kanugula, S.; Pegg, A. E.; Tainer, J. A. Active and Alkylated Human AGT Structures: A Novel Zinc Site, Inhibitor and Extrahelical Base Binding. *The EMBO Journal* **2000**, *19*, 1719–1730.

[79] Guengerich, F. P.; Fang, Q.; Liu, L.; Hachey, D. L.; Pegg, A. E. O6-Alkylguanine-DNA Alkyltransferase: Low pKa and High Reactivity of Cysteine 145. *Biochemistry* **2003**, *42*, 10965–10970.

[80] Pickersgill, R.; Harris, G.; Garman, E. Structure of Monoclinic Papain at 1.60 Å Resolution. *Acta Crystallogr., Sect. B: Struct. Sci* **1992**, *48*, 59–67.

[81] Barford, D.; Flint, A. J.; Tonks, N. K. Crystal Structure of Human Protein Tyrosine Phosphatase 1B. *Science* **1994**, *263*, 1397–1404.

[82] Pickersgill, R. W.; Rizkallah, P.; Harris, G. W.; Goodenough, P. W. Determination of the Structure of Papaya Protease Omega. *Acta Crystallogr., Sect. B: Struct. Sci* **1991**, *47*, 766–771.

[83] Stuckey, J. A.; Schubert, H. L.; Fauman, E. B.; Zhang, Z.-Y.; Dixon, J. E.; Saper, M. A. Crystal-Structure of Yersinia Protein-Tyrosine-Phosphatase at 2.5-Angstrom and the Complex with Tungstate. *Nature* **1994**, *370*, 571–575.

[84] Grimsley, G. R.; Scholtz, J. M.; Pace, C. N. A Summary of the Measured pK Values of the Ionizable Groups in Folded Proteins. *Protein Sci.* **2009**, *18*, 247–251.

[85] H++ (web-based computational prediction of protonation states and pK of ionizable groups in macromolecules). http://biophysics.cs.vt.edu/, Accessed: 2015-03-02.

[86] Warshel, A. Calculations of Enzymatic Reactions: Calculations of pKa, Proton Transfer Reactions, and General Acid Catalysis Reactions in Enzymes. *Biochemistry* **1981**, *20*, 3167–3177.

[87] Straatsma, T. P.; Berendsen, H. J. C. Free Energy of Ionic Hydration: Analysis of a Thermodynamic Integration Technique to Evaluate Free Energy Differences by Molecular Dynamics Simulations. *J. Chem. Phys.* **1988**, *89*, 5876–5886.

[88] Sugita, Y.; Okamoto, Y. Replica-Exchange Molecular Dynamics Method for Protein Folding. *Chem. Phys. Lett.* **1999**, *314*, 141–151.

[89] Sugita, Y.; Kitao, A.; Okamoto, Y. Multidimensional Replica-Exchange Method for Free-Energy Calculations. *J. Chem. Phys.* **2000**, *113*, 6042–6051.

[90] Fukunishi, H.; Watanabe, O.; Takada, S. On the Hamiltonian Replica Exchange Method for Efficient Sampling of Biomolecular Systems: Application to Protein Structure Prediction. *J. Chem. Phys.* **2002**, *116*, 9058–9067.

[91] Earl, D. J.; Deem, M. W. Parallel Tempering: Theory, Applications, and New Perspectives. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3910–3916.

[92] Meng, Y.; Sabri Dashti, D.; Roitberg, A. E. Computing Alchemical Free Energy Differences with Hamiltonian Replica Exchange Molecular Dynamics (H-REMD) Simulations. *J. Chem. Theory Comput.* **2011**, *7*, 2721–2727.

[93] Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79*, 926–935.

[94] Bussi, G.; Donadio, D.; Parrinello, M. Canonical Sampling Through Velocity Rescaling. *J. Chem. Phys.* **2007**, *126*, 014101.

[95] Parrinello, M.; Rahman, A. Polymorphic Transitions in Single Crystals: A New Molecular Dynamics Method. *J. Appl. Phys.* **1981**, *52*, 7182–7190.

[96] Nosé, S.; Klein, M. L. Constant Pressure Molecular Dynamics for Molecular Systems. *Mol. Phys.* **1983**, *50*, 1055–1076.

[97] Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. LINCS: A Linear Constraint Solver for Molecular Simulations. *J. Comput. Chem.* **1997**, *18*, 1463–1472.

[98] Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald: An N · log(N) method for Ewald sums in Large Systems. *J. Chem. Phys.* **1993**, *98*, 10089–10092.

[99] Pronk, S.; Páll, S.; Schulz, P., Roland Larsson; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D.; Hess, B.; Lindahl, E. GROMACS 4.5: A High-Throughput and Highly Parallel Open Source Molecular Simulation Toolkit. *Bioinformatics* **2013**, *29*, 845–854.

[100] Hub, J. S.; de Groot, B. L.; van der Spoel, D. g_wham – A Free Weighted Histogram Analysis Implementation Including Robust Error and Autocorrelation Estimates. *J. Chem. Theory. Comput.* **2010**, *6*, 3713–3720.

[101] Nicholls, A. Confidence limits, error bars and method comparison in molecular modeling. Part 1: The calculation of confidence intervals. *J. Comput.-Aided Mol. Des.* **2014**, *28*, 887–918.

[102] Beauchamp, K. A.; Lin, Y.-S.; Das, R.; Pande, V. S. Are Protein Force Fields Getting Better? A Systematic Benchmark on 524 Diverse NMR Measurements. *J. Chem. Theory Comput.* **2012**, *8*, 1409–1414.

[103] Cino, E. A.; Choy, W.-Y.; Karttunen, M. Comparison of Secondary Structure Formation Using 10 Different Force Fields in Microsecond Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **2012**, *8*, 2725–2740.

[104] Rauscher, S.; Gapsys, V.; Gajda, M. J.; Zweckstetter, M.; de Groot, B. L.; Grubmüller, H. Structural Ensembles of Intrinsically Disordered Proteins Depend Strongly on Force Field: A Comparison to Experiment. *J. Chem. Theory Comput.* **2015**, *11*, 5513–5524.

[105] Riahi, S.; Rowley, C. N. A Drude Polarizable Model for Liquid Hydrogen Sulfide. *J. Phys. Chem. B* **2013**, *117*, 5222–5229.

[106] Riahi, S.; Rowley, C. N. Solvation of Hydrogen Sulfide in Liquid Water and at the Water–Vapor Interface Using a Polarizable Force Field. *J. Phys. Chem. B* **2014**, *118*, 1373–1380.

[107] Riahi, S.; Rowley, C. N. Why Can Hydrogen Sulfide Permeate Cell Membranes? *J. Am. Chem. Soc.* **2014**, *136*, 15111–15113.

[108] Huang, J.; Lopes, P. E. M.; Roux, B.; Alexander D. MacKerell, J. Recent Advances in Polarizable Force Fields for Macromolecules: Microsecond Simulations of Proteins Using the Classical Drude Oscillator Model. *J. Phys. Chem. Lett.* **2014**, *5*, 3144–3150.

[109] Pearson, R. G. Hard and Soft Acids and Bases, HSAB, Part 1: Fundamental Principles. *J. Chem. Ed.* **1968**, *45*, 581.

[110] Crystal Structure Analysis of Oxidized Pseudomonas Aeruginosa Azurin at pH 5.5 and pH 9.0. *J. Mol. Biol.* **1991**, *221*, 765–772.

[111] Langella, E.; Improta, R.; Barone, V. Checking the pH-Induced Conformational Transition of Prion Protein by Molecular Dynamics Simulations: Effect of Protonation of Histidine Residues. *Biophys. J.* **2004**, *87*, 3623–3632.

[112] Lewis, S. D.; Johnson, F. A.; Shafer, J. A. Effect of Cysteine-25 on the Ionization of Histidine-159 in Papain As Determined by Proton Nuclear Magnetic Resonance Spectroscopy. Evidence for a Histidine-159-cysteine-25 Ion Pair and Its Possible Role in Catalysis. *Biochemistry* **1981**, *20*, 48–51.

[113] Mongan, J.; Case, D. A.; McCammon, J. A. Constant pH Molecular Dynamics in Generalized Born Implicit Solvent. *J. Comput. Chem* **2004**, *25*, 2038–2048.

[114] Swails, J. M.; York, D. M.; Roitberg, A. E. Constant pH Replica Exchange Molecular Dynamics in Explicit Solvent Using Discrete Protonation States: Implementation, Testing, and Validation. *J. Chem. Theory Comput.* **2014**, *10*, 1341–1352.

[115] Chen, Y.; Roux, B. Constant-pH Hybrid Nonequilibrium Molecular Dynamics–Monte Carlo Simulation Method. *J. Chem. Theory Comput.* **2015**, *11*, 3919–3931.

[116] Donnini, S.; Tegeler, F.; Groenhof, G.; Grubmüller, H. Constant pH Molecular Dynamics in Explicit Solvent with λ-Dynamics. *J. Chem. Theory Comput.* **2011**, *7*, 1962–1978.

[117] Donnini, S.; Ullmann, R. T.; Groenhof, G.; Grubmüller, H. Charge-Neutral Constant pH Molecular Dynamics Simulations Using a Parsimonious Proton Buffer. *J. Chem. Theory Comput.* **2016**, *12*, 1040–1051.

[118] Goh, G. B.; Hulbert, B. S.; Zhou, H.; Brooks, C. L. Constant pH Molecular Dynamics of Proteins in Explicit Solvent with Proton Tautomerism. *Proteins: Struct., Funct., Bioinf.* **2014**, *82*, 1319–1331.

[119] Lee, J.; Miller, B. T.; Damjanović, A.; Brooks, B. R. Constant pH Molecular Dynamics in Explicit Solvent with Enveloping Distribution Sampling and Hamiltonian Exchange. *J. Chem. Theory Comput.* **2014**, *10*, 2738–2750.

"A new scientific truth does not triumph by convincing its opponents and making them see the light, but rather because its opponents eventually die, and a new generation grows up that is familiar with it."

— Max Planck

3

The Hydration Structure of Methylthiolate from QM/MM Molecular Dynamics

This chapter is adapted with permission from: Awoonor-Williams, E. and Rowley, C. N. The Hydration Structure of Methylthiolate from QM/MM Molecular Dynamics *J. Chem. Phys.*, **2018**, 149, 045103-8.

Contents

3.1 Abstract

Thiols are widely present in biological systems, most notably as the side chain of cysteine amino acids in proteins. Thiols can be deprotonated to form a thiolate, which affords a diverse range of enzymatic activity and modes for chemical modification of proteins. Parameters for modeling thiolates using molecular mechanical force fields have not yet been validated, in part due to the lack of structural data on thiolate solvation. Here, the CHARMM36 and Amber models for thiolates in aqueous solutions are assessed using free energy perturbation and quantum mechanical/molecular mechanical (QM/MM) molecular dynamics (MD) simulations. The hydration structure of methylthiolate was calculated from 1 ns of QM/MM MD (PBE0-D3/def2-TZVP//TIP3P), which show that the water–S^- distances are approximately 2 Å with a coordination number near 6. The CHARMM thiolate parameters predict a thiolate S radius close to the QM/MM value and predict a hydration Gibbs energy of -329.2 kJ/mol, close to the experimental value of -318 kJ/mol. The cysteine thiolate model in the Amber force field underestimates the thiolate radius by ≈0.2 Å and overestimates the thiolate hydration energy by 119 kJ/mol because it uses the same Lennard-Jones parameters for thiolates as for thiols. A recent Drude polarizable model for methylthiolate with optimized thiolate parameters also performs well. SAPT2+ analysis indicates exchange repulsion is larger for the methylthiolate, consistent with it having a more diffuse electron density distribution in comparison to the parent thiol. These data demonstrate that it is important to define distinct nonbonded parameters for the protonated/deprotonated states of amino acid side chains in molecular mechanical force fields.

3.2 Introduction

Cysteine is unique among the amino acids due to its alkyl thiol side chain. Due to the weakness of the S–H bond, the acid dissociation constant (pK_a) of cysteine is moderate (∼8.6).[1] This allows cysteine to be deprotonated to form a thiolate anion under physiological conditions. Through this Brønsted acid mechanism, cysteine serves as a catalytic residue in some enzymes[2–6] and can undergo a broad range of post-translational modification reactions.[7–11]

Despite the importance of thiolate chemistry, the physical description of thiolates in solution lags behind those of other anions; we have been unable to find experimental studies of the hydration structure of thiolates that would establish basic quantities of the ion–water distance and coordination number. The hydration structure of other anions has been determined by a combination of experiment and computation. For example, Pluhařová et al. found that the molecular dynamics simulations that matched the neutron scattering profiles of aqueous solutions of LiCl most closely predicted a Cl^- coordination number of 7.7.[12] Computer simulations could help understand and predict reactions involving cysteine, although this type of simulation requires an accurate molecular mechanical model of the deprotonated thiolate state. This is particularly important for the calculation of the pK_a of a cysteine residue in a protein using free energy perturbation (FEP) or constant-pH MD.[13,14]

The popular Amber and CHARMM force fields include parameters for deprotonated cysteine residues. The parameters used for thiolates in the CHARMM force field were included in early revisions and have been propagated into modern versions.[15] The Amber 99 force field includes a deprotonated cysteine residue (CYM),[16] although the same Lennard-Jones parameters are used for the cysteine thiol sulfur and the deprotonated thiolate form. No formal validation has been reported for either of these models.

Ab initio molecular dynamics (AIMD) has been widely used to study the hydration structure of solutes.[17–21] These simulations can provide a first-principles representation of the solvent–ion structure. This is especially valuable for thiolates because there is little experimental data to describe their solution structure. QM/MM MD is a variant of AIMD where the solute and first coordination sphere of waters can be described using QM while the balance of the solvent is described using MM. Significantly, QM/MM MD can be performed using atom-centered basis sets, so hybrid density functional theory (DFT) functionals can be used efficiently. These hybrid functionals mitigate the issues associated with delocalization error in these systems, which can be serious for weakly bound anions like thiolates.[22,23]

To investigate the solvation structure of thiolates, we have performed QM/MM MD simulations of a model thiolate (methylthiolate, CH_3S^-) and its parent thiol (methylthiol, CH_3SH) for comparison. The results of the simulations are compared to those predicted by the CHARMM and Amber force fields to assess their accuracy.

Symmetry Adapted Perturbation Theory (SAPT) is used to investigate the origin of difference in the thiol and thiolate hydration structure. The hydration energies of the MM models are also calculated and compared to the experimental value. Lastly, the solvation structures of methylthiol and methylthiolate structures calculated using the newly-released Drude polarizable force field are compared to the QM/MM structure.

3.3 Computational Methods

3.3.1 QM/MM Simulations

A QM/MM model for methylthiolate (CH_3S^- (aq)) in a 14 Å sphere of water molecules was constructed with CHARMM c40b2[24] interfaced with TURBOMOLE 7.0[25] using the CHARMM-TURBOMOLE interface.[26] The QM region was comprised of the methylthiolate anion and 12 water molecules (Figure 3.1). These components were represented using the PBE0 exchange-correlation functional[27] and the def2-TZVP basis set.[28] This functional has been shown to describe the electrostatic moments and polarizability of molecules accurately[29] and the high exact-exchange component of functional mitigates the effect of delocalization error.[22] The D3 correction for dispersion was included.[30] The MM region was represented using 428 TIP3P-model water molecules.[31,32] In the CHARMM–TURBOMOLE QM/MM implementation, the QM atoms are polarized by the partial atomic charges of the MM region. The forces between the MM point charges and QM atoms are calculated rigorously through one-electron integrals between each MM atomic charge and the QM electron density as well as the Coulombic interaction between the QM nuclei and the MM point charges. Additionally, Lennard-Jones interactions are calculated between the QM region and the MM region, where the Lennard-Jones parameters of the CHARMM force field are used for the atoms in the QM region. The sulfur atom was restrained to the origin using a harmonic restraint (k=418 kJ/mol/Å^2). MM water molecules were restrained to remain within a 14 Å radius around the sulfur atom using a half-harmonic potential.

The general construction of QM/MM MD simulations of solutes in solution is for the solute and the water molecules closest to it to be described using the QM method. QM/MM simulations of ions in solution face the issue that the QM water molecules could diffuse away from the ion and be replaced by MM waters. A variety of boundary

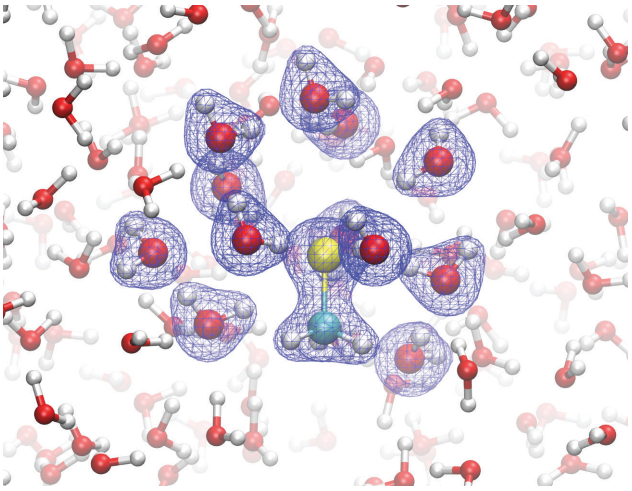

Figure 3.1: A representative snapshot of the QM/MM system. The electron density of the QM region, comprised of the methylthiolate and 12 nearest water molecules, is shown by blue mesh.

methods have been developed to address this. For example, the Adaptive Buffered Force method interpolates between a QM representation and an MM representation of a solvent molecule as it moves across a predefined buffer region.[33]

The Flexible Inner Region Ensemble Separator (FIRES) employs an alternative approach to dividing the QM and MM regions.[34] This method takes advantage of the separability of a configurational integral of the system, which can be rigorously rewritten as the product of integrals over the first m water molecules multiplied by the configurational integral over the remaining solvent molecules,

$$\begin{aligned} Z &= \int \mathrm{d}\mathbf{r}_{solute} \tfrac{1}{N!} \int \cdots \int \mathrm{d}\mathbf{r}_1 \cdots \mathrm{d}\mathbf{r}_N \exp\left(\tfrac{-\mathcal{V}}{k_B T}\right) \\ &= \tfrac{1}{m!} \int \mathrm{d}\mathbf{r}_{solute} \int \cdots \int \mathrm{d}\mathbf{r}_1 \cdots \mathrm{d}\mathbf{r}_m \\ &\times \tfrac{1}{(N-m)!} \int' \cdots \int' \mathrm{d}\mathbf{r}_{m+1} \cdots \mathrm{d}\mathbf{r}_N \exp\left(\tfrac{-\mathcal{V}}{k_B T}\right) \end{aligned} \tag{3.1}$$

here $\int'$ denotes that the configurational integral only integrates the region of configurational space where $r > \max(r_1, r_2, \ldots, r_m)$ (i.e., those more distant from the ion than the outermost solute in the set of solute molecules with indices ranging from 1 to m).

In a QM/MM system, this separation can be used to define the m solvent molecules closest to the solute as the QM set, while those further from the solute are represented

using MM. This condition must be enforced in order to sample this distribution in an MD simulation. This can be achieved in an approximate manner by introducing a half-harmonic potential. At each time step, if the outermost QM water is further from the solute than the nearest MM water, a half-harmonic potential is applied to push the QM water towards the solute and push the MM water outward.

$$\mathcal{V}_{FIRES} = k_{FIRES} \left(r_k - R_{inner} \right)^2 , \text{if } r_k < R_{inner} \tag{3.2}$$

where $R_{inner} = \max(r_1, r_2, \ldots, r_m)$ and k is the index of the MM solvent molecule that has violated the FIRES criterion by moving closer to the solute than the outermost QM solvent molecule and r_k is the distance between this molecule and the solute.

In this study, the FIRES boundary potential was used to restrict the QM water molecules to remain closest to the sulfur atom throughout the simulation. The CHARMM and TURBOMOLE input files for these simulations are included in the supplementary information of Ref. 35. Three simulations were initiated from different snapshots of an equilibrated MM simulation. The first 50 ps of each simulation was discarded prior to a 334 ps sampling simulation. The QM/MM radial distribution functions (RDFs) were calculated from these three trajectories, which totaled 1 ns of MD.

3.3.2 Molecular Mechanical Simulations

The pure MM MD simulations were performed using NAMD 2.12.[36] A $32 \times 32 \times 32$ Å simulation cell was constructed, which contained the solute and 987 TIP3P-model water molecules. The MM RDFs were calculated from a 40 ns simulation. The simulations sampled an isothermal-isobaric ensemble with a temperature of 298 K and a pressure 101.325 kPa using a Langevin thermostat with a damping frequency of 1 ps^{-1} and a Nosé-Hoover Langevin piston barostat with a period of 2 ps.

The Gibbs energies of hydration were calculated using CHARMM c40b1 with the thermodynamic integration technique. To calculate the relative solvation energy of methylthiol and methylthiolate, the charges and Lennard-Jones parameters for methylthiol were simultaneously transformed to those for methylthiolate using thermodynamic integration, where values of λ were calculated at 11-evenly spaced intervals between [0, 1]. Each value of λ was run for a 1 ns equilibration simulation, followed

by a 2 ns sampling simulation. The solvation energy of methylthiol for the CHARMM and Amber models ($\Delta G_{hydr.}$(MeSH) = −0.04 and −1.12 kJ/mol, respectively) from Ref. (37) were used to calculate the absolute solvation energy of methylthiolate from the relative solvation energy of methylthiol and methylthiolate. Tables 3.1 and 3.2 lists the Lennard-Jones parameters for methylthiol and methylthiolate as adopted by the Amber and CHARMM force field models, following the definition of the Lennard-Jones potential of,

$$\mathcal{V}_{LJ} = 4\epsilon \left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6} \right] \tag{3.3}$$

Gibbs energies were calculated from the TI time series using the Weighted Histogram Analysis Method (WHAM). The calculated hydration energies corrected for an interfacial potential of -520 mV, which corresponds to a correction of 48.2 kJ/mol.[38] Uncertainties were estimated by dividing the production trajectory into three equal segments, calculating the hydration energy from this segment, and then taking the standard deviation of the three Gibbs energies.

Table 3.1: Force field parameters for methylthiol.

Atom		Amber	CHARMM
S	ϵ (kJ/mol)	1.05	1.88
	σ (Å)	3.56	3.56
	q (e)	-0.31	-0.23
C	ϵ (kJ/mol)	0.46	0.46
	σ (Å)	3.40	3.65
	q (e)	-0.12	-0.20

Table 3.2: Force field parameters for methylthiolate.

Atom		Amber	CHARMM
S^-	ϵ (kJ/mol)	1.05	1.97
	σ (Å)	3.56	3.92
	q (e)	-0.88	-0.80
C	ϵ (kJ/mol)	0.46	0.46
	σ (Å)	3.40	3.92
	q (e)	-0.24	-0.47

3.3.3 Symmetry Adapted Perturbation Theory

Methylthiol/methylthiolate–water potential energy surfaces (PES) were calculated using the SAPT2+ method[39–41] with the aug-cc-pVTZ basis set and an auxiliary, density-fitting basis set. All SAPT2+ calculations were performed using version 1.1 of the Psi4 code.[42] The water and methylthiol/methylthiolate molecules were optimized in isolation using MP2/aug-cc-pVTZ. For each position on the PES, an energy minimization was performed using MP2/aug-cc-pVTZ where ∠C–S—O angle was allowed to change in a constrained energy minimization calculation but the remaining structural degrees of freedom were fixed. Using these structures, SAPT2+ energies were calculated on a grid spanning sulfur–oxygen distances between 2.5 and 5.0 Å, with a grid spacing of 0.1 Å and ∠ S—O–H angles spanning 0°–180° at 5° increments. Orientationally-averaged potential energies were calculated by numerically integrating over the θ angles for each value of r (Eqn. 3.4), where each configuration was weighted according to the Boltzmann distribution at 298 K (Eqn. 3.4).[43]

$$\langle \mathcal{V}(r) \rangle = \frac{\int_0^\pi \mathcal{V}(r,\theta) \exp\left(\frac{-\mathcal{V}(r,\theta)}{k_B \mathrm{T}}\right) \sin\theta \mathrm{d}\theta}{\int_0^\pi \exp\left(\frac{-\mathcal{V}(r,\theta)}{k_B \mathrm{T}}\right) \sin\theta \mathrm{d}\theta} \tag{3.4}$$

Here, $\mathcal{V}(r,\theta)$ is the interaction energy between the water and methylthiol/methylthiolate at a given value of r (the sulfur–oxygen distance) and θ (the sulfur—oxygen–hydrogen angle).

3.4 Results and Discussion

3.4.1 Radial Distribution Functions

Methylthiol

All three models predict similar hydration structures for methylthiol, which has a broad first peak of the S—O(H_2) RDF centered at 3.6 Å (Figure 3.2 (a)). The first peak is somewhat higher in the QM/MM and CHARMM models, corresponding to an increased hydration number of approximately 13 in the QM/MM model vs approximately 10–11 in the MM models (Table 3.3). The S—HOH RDF shows only

a small shoulder on the left side of the first peak for all three models (Figure 3.2 (b)). This is consistent with previous studies that concluded that thiol sulfurs are poor hydrogen bond acceptors due to their large radius and modest electronegativity.[20,44] These RDF's are consistent with the thiols being hydrophobic solutes, with very limited hydrogen bonding interactions with the aqueous solvent.

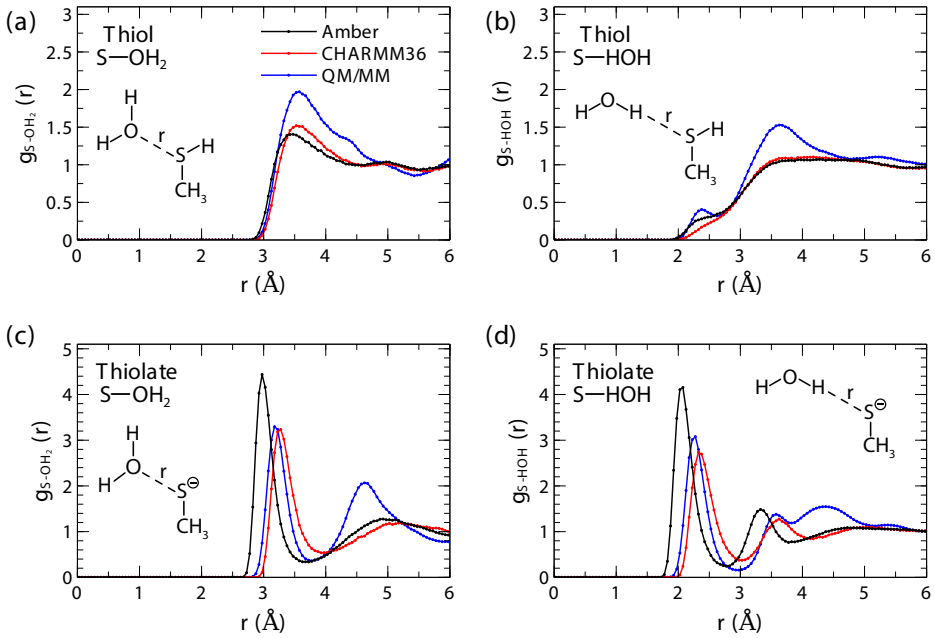

Figure 3.2: Radial distribution functions of methylthiol (upper) and methylthiolate (lower) models in aqueous solution using Amber, CHARMM, and QM/MM models. **Thiol**: (a) **S**—**O**(H_2), (b) **S**—**H**OH; **Thiolate**: (c) **S**—**O**(H_2), and (d) **S**—**H**OH. Note: The chemical structures are schematics to indicate the geometric variables. Figure 3.1 provides a more representative rendering of the solvation structure.

Table 3.3: Coordination numbers of methylthiol and methylthiolate from simulations using the CHARMM, Amber, and QM/MM models. The thiolate coordination numbers are calculated by integrating $g(r)$ between zero and the first minimum. The thiol coordination numbers are calculated by integrating over the interval, r=[0, 4.5 Å].

Models	Thiol	Thiolate
CHARMM	10.7	7.2
Amber	10.5	6.7
Drude	9.3	7.1
QM/MM	13.5	6.0

Methylthiolate

The QM/MM RDF's of methylthiolate are typical of a diffuse anion. There is a high but broad first peak in the S—O(H_2) RDF that corresponds to 6 water molecules in a disordered first coordination sphere (Figure 3.2 (c), Table 3.3). This ionic radius and coordination number is similar to those reported for the Cl^- anion ($g(r)$ is maximum at $r = 3.15$ Å, $n_c = 6 - 6.5$).[45,46] The typical coordination mode is for the water molecules to donate a hydrogen bond to the thiolate sulfur, resulting in the peak observed in the S—HOH RDF at $r = 2.3$ Å (Figure 3.2 (d)).

There is good agreement between the RDF's calculated using the CHARMM force field and the QM/MM simulations; the maxima of the S—O(H_2) and S—HOH peaks for the CHARMM36 model occur at similar positions as in the QM/MM RDFs. In comparison, the location of the first peaks of the S—O(H_2) and S—HOH RDF's calculated using the Amber model occur at significantly smaller radii. This trend is also apparent in optimized structures of a single methylthiolate–water pair; the Amber S—H distance is only 2.02 Å while the CHARMM S—H distance is 2.47 Å.

The difference between the CHARMM and Amber force fields can be attributed to the Lennard-Jones parameters of the thiolate sulfur atom. The Amber model for thiolates only changes the atomic charges of the cysteine thiol residue; the Lennard-Jones parameters are the same as those used for the thiol (Table 3.1, 3.2). As an anion, the ionic radius of S^- is considerably larger than that of neutral S. A realistic description of thiolates requires distinct Lennard-Jones parameters from the parent thiol to reflect this more diffuse electron density distribution.

The second coordination sphere of the QM/MM model had a higher maximum than the MM models and the S–HOH RDF shows a splitting in the peak of the second coordination sphere. As the number of atoms described using QM must be kept low in order to achieve long-time scale simulations using accurate QM/MM methods, part of the 2nd coordination sphere is comprised of MM waters, which will indirectly affect the first coordination sphere. The interactions between two QM waters tends to be stronger than between a QM water and an MM water due to polarization and MM sphere with molecules in the first coordination sphere than they have with MM waters. Inevitably, there is a transition between the QM representation and the MM representation in QM/MM calculations. When using the FIRES boundary, this will manifest itself over a range of radii that are within the QM region at some steps but in

the MM region in other steps. In principle, a larger set of QM waters could be defined so that the entire 2nd coordination sphere is always in the QM region, although the increased computational cost associated with this would limit the accuracy of the simulations. Alternatively, this could be attenuated by a larger QM region or by pair-specific Lennard-Jones terms between the QM and the MM water molecules that would make the interactions between QM and MM waters closer in strength.

3.4.2 Hydration Energies

The absolute hydration energies of methylthiolate calculated using the CHARMM and Amber force fields are presented in Table 3.4. The experimental solvation energy was taken from Ref. 47, which was estimated from the experimental methylthiolate pK_a of 10.30. The hydration energy calculated using the CHARMM model is in remarkable agreement with the experimental value, differing only by 11 kJ/mol. In contrast, the Amber model overestimates the hydration energy by 119 kJ/mol. This appears to reflect the use of the same Lennard-Jones parameters for the thiol and thiolate states in the Amber force field, which results in the anomalously short S^-–water distances seen in the RDF analysis.

Table 3.4: Absolute hydration free energies of methylthiolate for molecular mechanical models.

Models	ΔG (kJ/mol)
CHARMM	-329.2 ± 0.3
Amber	-437.0 ± 0.8
Drude[1]	-308.4
exptl. [2]	-318

[1] Ref. 48
[2] Ref. 47

Symmetry Adapted Perturbation Theory

The differences in the thiol and thiolate water RDF's can be understood using Symmetry Adapted Perturbation Theory (SAPT) analysis. The total potential energy and its components for the interaction between methylthiolate/methylthiol and a water molecule calculated using SAPT2+ are plotted in Figure 3.3. The total potential energy of interaction is much stronger between methylthiolate and water than methylthiol and water (−62 kJ/mol vs. −10 kJ/mol). The SAPT2+ methylthiolate–water interaction energy is in good agreement with a benchmark CCSD(T)/aug-cc-pVTZ calculation (Table 3.5), so we can be confident in the SAPT2+ analysis.

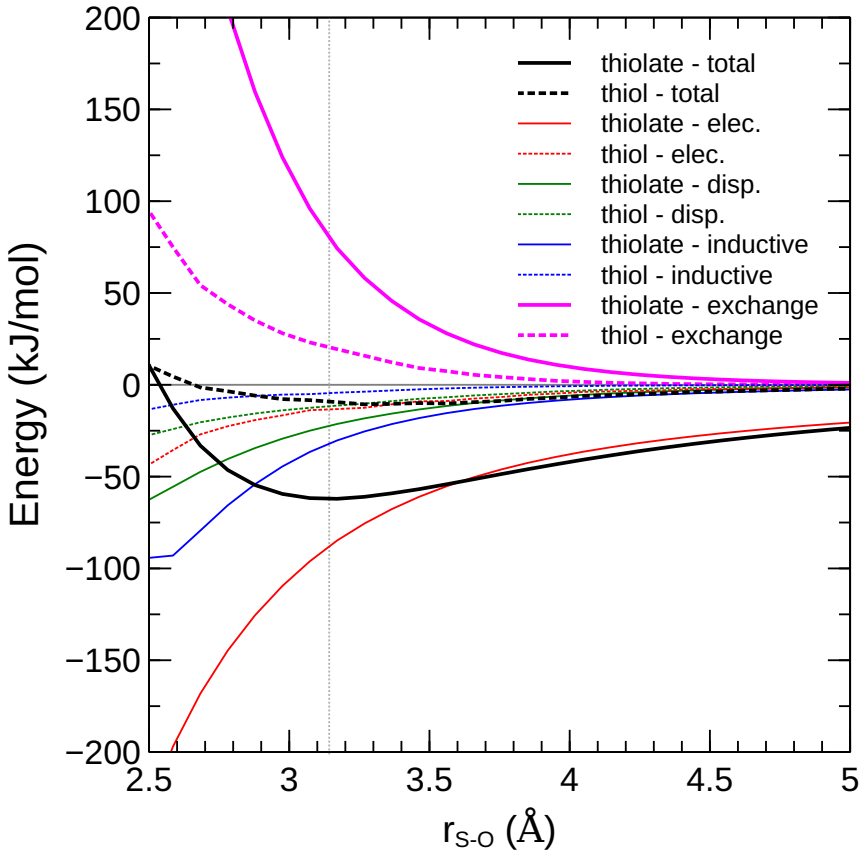

Figure 3.3: Rotationally averaged SAPT2+ potential energy surfaces for the interaction between methylthiol/methylthiolate with a water molecule. The position of the methylthiolate–water potential energy minimum is indicated by the vertical dotted line. The solid lines correspond to the methylthiolate–water interaction, while the dashed lines represent the methylthiol–water interactions.

Table 3.5: Water–methylthiolate interaction energies and S^-—H–OH distance in the minimum-energy structure.

Models	$\Delta E_{\text{complex}}$ (kJ/mol)	r_{S-H} (Å)
CHARMM	-55.0	2.47
Amber	-69.9	2.02
Drude	-58.0	2.19
SAPT2+	-62.1	2.21
CCSD(T)/aug-cc-pVTZ//MP2/aug-cc-pVTZ	-64.7	2.18

The stronger interaction between the methylthiolate and water in comparison to the interaction between methylthiol and water largely results from stronger electrostatic interactions between the anionic methylthiolate and the water molecule in comparison to the dipole–dipole interactions of the weakly-polar methylthiol and water. The methylthiolate–water interaction also has a strong inductive component due to the exceptionally high polarizability of the methylthiolate and the increased strength of the polarizing electric field. Likewise, the methylthiolate–water dispersion interaction is also stronger. In opposition to these stronger methylthiolate–water interactions, the methylthiolate–water exchange repulsion is also considerably stronger than the methylthiol–water exchange repulsion. This is consistent with the atomic radius of the sulfur being effectively larger for the thiolate, as the electron density cloud of the sulfur anion is more diffuse than the neutral thiol. These interactions are consistent with the thiolate having S—O distances that are incrementally shorter than for the thiol in the QM/MM RDF's. It also highlights the importance of defining different Lennard-Jones parameters for an ion and its parent molecule, as the underlying exchange-repulsion and dispersion interactions can be considerably different.

Neither of the molecular mechanical models are in good agreement with the SAPT2+ energy. The CHARMM model underestimates the interaction energy but overestimates the r_{S-H} distance, while the Amber model overestimates the interaction energy but underestimates the r_{S-H} distance (Table 3.5). The short Amber distance is consistent with the short S^-—HOH distances in the calculated RDF. It should be noted that the water–ion interaction energies of non-polarizable molecular mechanical models typically do not match QM or experimental interaction energies closely because they are generally parameterized to reproduce ion solvation Gibbs energies.[49] Because of the approximate nature of non-polarizable force fields, parameters that are

effective for calculating interaction energies are not generally effective at describing properties like hydration energies. Nevertheless, the SAPT2+ analysis demonstrates that the protonated and deprotonated states of the thiol have intermolecular interactions with water that are fundamentally different, which should be accounted for by the non-bonded parameters of the two forms.

Protein pK_a Calculations

The CHARMM36 and Amber force fields have both been used to calculate the pK_a's of cysteine residues in proteins. For a test set consisting of 18 cysteine residues in 12 proteins, the CHARMM36 force field was found to predict cysteine pK_a's more accurately than the Amber force field when using the replica-exchange molecular dynamics thermodynamic integration (REMD-TI) method.[13] For example, Cys232 of α-1-antitrypsin has a reported pK_a of 6.86 ± 0.05.[50] REMD-TI pK_a calculations using the CHARMM36 force field are fairly effective at predicting this cysteine pK_a value (pK_a=7.6 ± 0.4), while simulations using Amber force field significantly overestimate it (pK_a=9.4 ± 0.7).[13] Analysis of Cys232–S^-—water RDFs from 12-ns MD simulation of the thiolate state of this protein in TIP3P model water show similar trends as previously observed (Figure 3.4). A representative configuration of Cys232 thiolate in α-1-antitrypsin and water molecules within 5 Å of the thiolate is highlighted in Figure 3.4 (a). The thiolate S–water distances computed using the CHARMM36 force field is slightly larger (≈0.3 Å) than that of the Amber force field (Figure 3.4 (b), (c)). Consequently, the location of the first peaks of the S—O(H_2) and S—HOH RDFs occur at significantly shorter distances for the Amber model than for the CHARMM model. Protein pK_a simulations are generally relative predictions, so accurate pK_a predictions are still possible even if the titratable residue–water interactions are imperfect. Nevertheless, this analysis shows that the water–thiolate interactions of cysteine residues in proteins are strongly affected by the Lennard-Jones parameters of the sulfur thiolate.

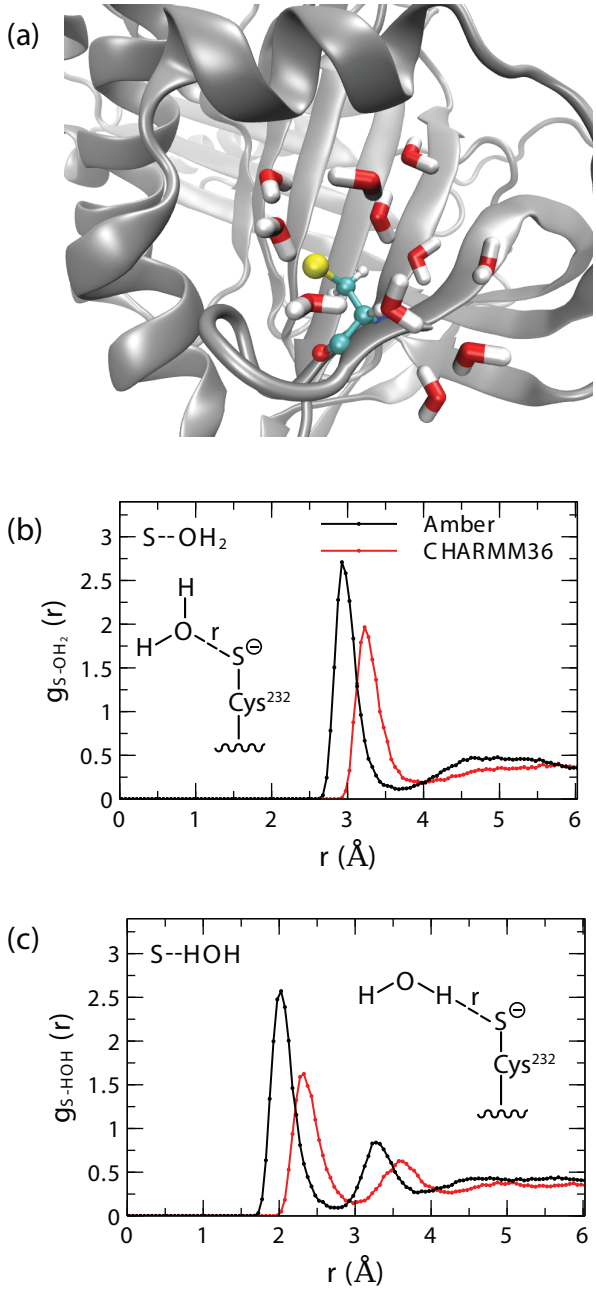

Figure 3.4: Representative configuration and radial distribution plots of Cys232 in α-1-antitrypsin. (a) Explicit solvent representation of Cys232 thiolate in α-1-antitrypsin (PDB ID: 1QLP); Radial distribution function of Cys232 thiolate with solvent water molecules, (b) **S**—**O**(H_2) and (c) **S**—**H**OH.

3.4.3 Drude Polarizable Force Field

A force field based on the classical Drude oscillator model[51] has recently been developed for simulations of aqueous methythiol and methylthiolate.[48] The methylthiolate model was parameterized to provide the correct hydration energy when used with the SWM4-NDP polarizable water model[52] and is in good agreement with the experimental hydration energy. To evaluate this model, we have plotted its RDF's against those calculated from our QM/MM simulations (Figure 3.5). The Drude model RDFs agree reasonably well with the QM/MM RDFs. Both models yield very similar values for the location of the first peaks of the S—O(H_2) and S—HOH distances, particularly for methylthiolate. The RDF peaks of methylthiol are, however, higher for the QM/MM model than the Drude model and the Drude model coordination number of methylthiol is 9.3, considerably lower than the QM/MM value.

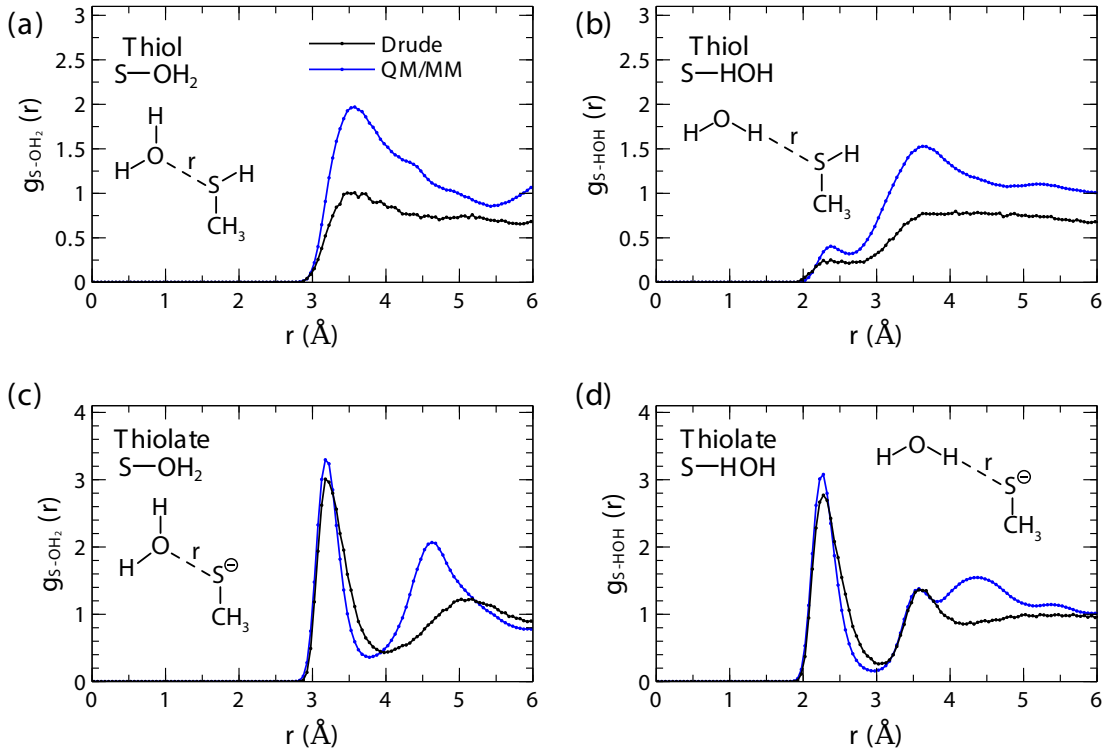

Figure 3.5: Radial distribution functions of methylthiol (upper) and methylthiolate (lower) models in aqueous solution using the classical Drude oscillator and QM/MM models. **Thiol**: (a) **S**—**O**(H_2), (b) **S**—**H**OH; **Thiolate**: (c) **S**—**O**(H_2), and (d) **S**—**H**OH.

The methylthiolate–water complexation energy and structure are also improved. The potential-energy-minimum S—H–OH distance for the Drude model is 2.17 Å, which is very close to QM value of 2.18 Å. Based on these results, these parameters are expected to provide reasonably-accurate descriptions of aqueous thiolates that could provide improved accuracy over the non-polarizable MM models when polarization effects are significant.

3.5 Conclusions

QM/MM MD simulations were used to characterize the solution structure of methylthiolate. These simulations and free energy perturbation calculations were used to evaluate the models for methylthiolate for Amber and CHARMM36 protein force fields. For comparison, simulations were also performed on methylthiol.

The calculated coordination structure of the thiolate is characteristic of a large anion. The solvation structure of the thiolate is more rigid than the corresponding thiol, consistent with a transition from a hydrophobic solute to an anionic solute. SAPT analysis showed that although the charge–dipole attraction between the thiolate and the water molecules are much stronger than the dipole–dipole interactions between the water and the thiols, the sulfur–water distance only decreases by ≈0.2 Å because the exchange repulsion is more significant at larger distances for the more diffuse thiolate sulfur.

The CHARMM model for the thiolate is in better agreement with the QM/MM RDF and experimental hydration energy than the Amber model. This can be attributed to the Lennard-Jones parameters used for the thiolate sulfur. The CHARMM force field uses Lennard-Jones parameters with a larger radius for the thiolate sulfur, while the Amber force field uses the same parameters for both the thiol and the thiolate. The QM/MM and FEP data indicate that distinct non-bonded parameters must be defined for the ionic forms of molecules. This is significant for the calculation of the pK_a's of titratable residues of proteins because the same Lennard-Jones parameters are often used for all protonation states. The development of optimized parameters could improve the accuracy of protein pK_a calculations, which remains a major challenge.

At present, the thiolate parameters included in the CHARMM36 force field provide

a generally good description of the solvation structure and energies of methylthiolate. The Drude polarizable force field parameters of Lin et al.[48] were also found to be effective for describing the solution structure of methylthiol and methylthiolate, providing a model that incorporates the effect of induced polarization. Further validation and development will be necessary to show that these models are generally reliable in non-aqueous environments.

Bibliography

[1] Thurlkill, R. L.; Grimsley, G. R.; Scholtz, J. M.; Pace, C. N. pK Values of the Ionizable Groups of Proteins. *Protein Sci.* **2006**, *15*, 1214–1218.

[2] Lewis, S. D.; Johnson, F. A.; Shafer, J. A. Effect of Cysteine-25 on the Ionization of Histidine-159 in Papain As Determined by Proton Nuclear Magnetic Resonance Spectroscopy. Evidence for a Histidine-159-cysteine-25 Ion Pair and Its Possible Role in Catalysis. *Biochemistry* **1981**, *20*, 48–51.

[3] Otto, H.-H.; Schirmeister, T. Cysteine Proteases and Their Inhibitors. *Chem. Rev.* **1997**, *97*, 133–172.

[4] Denu, J. M.; Dixon, J. E. Protein Tyrosine Phosphatases: Mechanisms of Catalysis and Regulation. *Curr. Opin. Chem. Biol.* **1998**, *2*, 633–641.

[5] Bartlett, G. J.; Porter, C. T.; Borkakoti, N.; Thornton, J. M. Analysis of catalytic residues in enzyme active sites. *J. Mol. Biol.* **2002**, *324*, 105–121.

[6] Giles, N. M.; Giles, G. I.; Jacob, C. Multiple roles of cysteine in biocatalysis. *Biochem. Biophys. Res. Commun.* **2003**, *300*, 1–4.

[7] Smith, J. M.; Rowley, C. N. Automated computational screening of the thiol reactivity of substituted alkenes. *J. Comput.-Aided Mol. Des.* **2015**, *29*, 725–735.

[8] deGruyter, J. N.; Malins, L. R.; Baran, P. S. Residue-Specific Peptide Modification: A Chemist's Guide. *Biochemistry* **2017**, *56*, 3863–3873.

[9] Chauvin, J.-P. R.; Griesser, M.; Pratt, D. A. Hydropersulfides: H-Atom Transfer Agents Par Excellence. *J. Am. Chem. Soc.* **2017**, *139*, 6484–6493.

[10] Chauvin, J.-P. R.; Pratt, D. A. On the Reactions of Thiols, Sulfenic Acids, and Sulfinic Acids with Hydrogen Peroxide. *Angew. Chem. Int. Ed* **2017**, *56*, 6255–6259.

[11] Awoonor-Williams, E.; Walsh, A. G.; Rowley, C. N. Modeling covalent-modifier drugs. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* **2017**, *1865*, 1664 – 1675, Biophysics in Canada.

[12] Pluhařová, E.; Fischer, H. E.; Mason, P. E.; Jungwirth, P. Hydration of the chloride ion in concentrated aqueous solutions using neutron scattering and molecular dynamics. *Mol. Phys.* **2014**, *112*, 1230–1240.

[13] Awoonor-Williams, E.; Rowley, C. N. Evaluation of Methods for the Calculation of the pKa of Cysteine Residues in Proteins. *J. Chem. Theory Comput.* **2016**, *12*, 4662–4673.

[14] Radak, B. K.; Chipot, C.; Suh, D.; Jo, S.; Jiang, W.; Phillips, J. C.; Schulten, K.; Roux, B. Constant-pH Molecular Dynamics Simulations for Large Biomolecular Systems. *J. Chem. Theory Comput.* **2017**, *13*, 5933–5944.

[15] Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E. M.; Mittal, J.; Feig, M.; Alexander D. MacKerell, J. Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone ϕ, ψ and Side-Chain χ1 and χ2 Dihedral Angles. *J. Chem. Theory Comput.* **2012**, *8*, 3257–3273.

[16] 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*, 1049–1074.

[17] Lyubartsev, A.; Laasonen, K.; Laaksonen, A. d. Hydration of Li+ ion. An ab initio molecular dynamics simulation. *J. Chem. Phys.* **2001**, *114*, 3120–3126.

[18] Armunanto, R.; Schwenk, C. F.; Tran, H. T.; Rode, B. M. Structure and dynamics of Au+ ion in aqueous solution: ab initio QM/MM MD simulations. *J. Am. Chem. Soc.* **2004**, *126*, 2582–2587.

[19] Riahi, S.; Roux, B.; Rowley, C. N. QM/MM molecular dynamics simulations of the hydration of Mg(II) and Zn(II) ions. *Can J. Chem.* **2013**, *91*, 552–558.

[20] Riahi, S.; Rowley, C. N. Solvation of Hydrogen Sulfide in Liquid Water and at the Water–Vapor Interface Using a Polarizable Force Field. *J. Phys. Chem. B* **2014**, *118*, 1373–1380.

[21] Awoonor-Williams, E.; Rowley, C. N. The hydration structure of carbon monoxide by ab initio methods. *J. Chem. Phys.* **2017**, *146*, 034503.

[22] Smith, J. M.; Jami Alahmadi, Y.; Rowley, C. N. Range-Separated DFT Functionals are Necessary to Model Thio-Michael Additions. *J. Chem. Theory Comput.* **2013**, *9*, 4860–4865.

[23] Kim, M.-C.; Sim, E.; Burke, K. Understanding and Reducing Errors in Density Functional Calculations. *Phys. Rev. Lett.* **2013**, *111*, 073003.

[24] Brooks, B. R.; Brooks, I. C. L.; Mackerell, J. A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; et al., CHARMM: The Biomolecular Simulation Program. *J. Comput. Chem.* **2009**, *30*, 1545–1614.

[25] TURBOMOLE V7.0 2015, a development of University of Karlsruhe and Forschungszentrum Karlsruhe GmbH, 1989-2007, TURBOMOLE GmbH, since 2007; available from
`http://www.turbomole.com`.

[26] Riahi, S.; Rowley, C. N. The CHARMM–TURBOMOLE interface for efficient and accurate QM/MM molecular dynamics, free energies, and excited state properties. *J. Comput. Chem.* **2014**, *35*, 2076–2086.

[27] Adamo, C.; Barone, V. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *J. Chem. Phys.* **1999**, *110*, 6158–6170.

[28] Weigend, F.; Ahlrichs, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297–3305.

[29] Hickey, A. L.; Rowley, C. N. Benchmarking Quantum Chemical Methods for the Calculation of Molecular Dipole Moments and Polarizabilities. *J. Phys. Chem. A* **2014**, *118*, 3678–3687.

[30] Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.* **2010**, *132*.

[31] Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79*, 926–935.

[32] A. D. MacKerell, J. et al. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.

[33] Bernstein, N.; Varnai, C.; Solt, I.; Winfield, S. A.; Payne, M. C.; Simon, I.; Fuxreiter, M.; Csanyi, G. QM/MM simulation of liquid water with an adaptive quantum region. *Phys. Chem. Chem. Phys.* **2012**, *14*, 646–656.

[34] Rowley, C. N.; Roux, B. The Solvation Structure of Na^+ and K^+ in Liquid Water Determined from High Level ab Initio Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **2012**, *8*, 3526–3535.

[35] Awoonor-Williams, E.; Rowley, C. N. The hydration structure of methylthiolate from QM/MM molecular dynamics. *J. Chem. Phys.* **2018**, *149*, 045103.

[36] Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kalé, L.; Schulten, K. Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **2005**, *26*, 1781–1802.

[37] Walters, E. T.; Mohebifar, M.; Johnson, E. R.; Rowley, C. N. Evaluating the London Dispersion Coefficients of Protein Force Fields Using the Exchange-Hole Dipole Moment Model. *J. Phys. Chem. B* **2018**, *122*, 6690–6701.

[38] Harder, E.; Roux, B. On the origin of the electrostatic potential difference at a liquid-vacuum interface. *J. Chem. Phys.* **2008**, *129*, 234706.

[39] Hohenstein, E. G.; Sherrill, C. D. Density fitting of intramonomer correlation effects in symmetry-adapted perturbation theory. *J. Chem. Phys.* **2010**, *133*, 014101.

[40] Hohenstein, E. G.; Sherrill, C. D. Wavefunction methods for noncovalent interactions. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2012**, *2*, 304–326.

[41] Parker, T. M.; Burns, L. A.; Parrish, R. M.; Ryno, A. G.; Sherrill, C. D. Levels of symmetry adapted perturbation theory (SAPT). I. Efficiency and performance for interaction energies. *J. Chem. Phys.* **2014**, *140*, 094106.

[42] Parrish, R. M. et al. Psi4 1.1: An Open-Source Electronic Structure Program Emphasizing Automation, Advanced Libraries, and Interoperability. *J. Chem. Theory Comput.* **2017**, *13*, 3185–3197.

[43] Dill, K.; Bromberg, S.; Stigter, D. *Molecular Driving Forces: Statistical Thermodynamics in Chemistry and Biology*; Garland Science, 2003.

[44] Riahi, S.; Rowley, C. N. A Drude Polarizable Model for Liquid Hydrogen Sulfide. *J. Phys. Chem. B* **2013**, *117*, 5222–5229.

[45] Zhang, C.; Pham, T. A.; Gygi, F.; Galli, G. Communication: Electronic structure of the solvated chloride anion from first principles molecular dynamics. *J. Chem. Phys.* **2013**, *138*, 181102.

[46] Lamoureux, G.; Roux, B. Absolute hydration free energy scale for alkali and halide ions established from simulations with a polarizable force field. *J. Phys. Chem. B* **2006**, *110*, 3308–3322.

[47] Sitkoff, D.; Sharp, K. A.; Honig, B. Accurate Calculation of Hydration Free Energies Using Macroscopic Solvent Models. *J. Phys. Chem.* **1994**, *98*, 1978–1988.

[48] Lin, F.-Y.; Lopes, P. E. M.; Harder, E.; Roux, B.; MacKerell, A. D. Polarizable Force Field for Molecular Ions Based on the Classical Drude Oscillator. *J. Chem. Inf. Model.* **2018**, *58*, 993–1004.

[49] Whitfield, T. W.; Varma, S.; Harder, E.; Lamoureux, G.; Rempe, S. B.; Roux, B. Theoretical Study of Aqueous Solvation of K+ Comparing ab Initio, Polarizable, and Fixed-Charge Models. *J. Chem. Theory Comput.* **2007**, *3*, 2068–2082.

[50] Griffiths, S. W.; King, J.; Cooney, C. L. The reactivity and oxidation pathway of cysteine 232 in recombinant human α1-antitrypsin. *J. Biol. Chem.* **2002**, *277*, 25486–25492.

[51] Lamoureux, G.; Roux, B. Modeling induced polarization with classical Drude oscillators: Theory and molecular dynamics simulation algorithm. *J. Chem. Phys.* **2003**, *119*, 3025–3039.

[52] Lamoureux, G.; Harder, E.; Vorobyov, I. V.; Roux, B.; MacKerell Jr, A. D. A polarizable model of water for molecular dynamics simulations of biomolecules. *Chem. Phys. Lett.* **2006**, *418*, 245–249.

"I would rather have questions that can't be answered than answers that can't be questioned."

— Richard Feynman

4

How Reactive are Druggable Cysteines in Protein Kinases?

This chapter is adapted with permission from: Awoonor-Williams, E. and Rowley, C. N. How Reactive are Druggable Cysteines in Protein Kinases? *J. Chem. Inf. Model*, **2018**, 58(9), 1935–1946.

Contents

4.1 Abstract

Targeted covalent-inhibitors (TCIs) have been successfully developed as high-affinity and selective inhibitors of enzymes of the protein kinase family. These drugs typically act by undergoing an electrophilic addition with an active site cysteine residue, so design of a TCI begins with the identification of a "druggable" cysteine. These electrophilic additions generally require the deprotonation of the thiol to form a reactive anionic thiolate, so the acidity of the residue is a critical factor. Few experimental measurements of the pK_a's of druggable cysteines have been reported, so computational prediction could prove to be very important in selecting reactive cysteine targets. Here we report the computed pK_a's of druggable cysteines in select protein kinases which are of clinical relevance for targeted therapies. The pK_a's of the cysteines were calculated using advanced computational methods based on all-atom replica-exchange thermodynamic integration molecular dynamics simulations in explicit solvent. We found that the acidities of druggable cysteines within protein kinases are diverse and elevated, indicating enormous differences in their reactivity. Constant-pH molecular dynamics simulations were also performed on select protein kinases, with results confirming this varied range in the acidities of druggable cysteines. Many of these active-site cysteines have low exposure to solvent molecules, elevating their pK_a. Electrostatic interactions with nearby anionic residues also elevate the pK_a's of cysteine residues in the active site. The results suggest that some cysteine residues within kinase binding sites will be slow to react with a TCI because of their low acidity. Several oncogenic kinase mutations were also modelled and found to have pK_a's similar to that of the wild-type kinase.

4.2 Introduction

The family of human protein kinases contains over 500 members,[1] which play critical biological roles in cellular processes such as signal transduction, cell metabolism, and apoptosis. Dysregulation, overexpression, and mutation in protein kinases has been linked to many proliferative human diseases and disorders, notably cancer and inflammation.[2] Kinases have proven to be excellent drug targets;[3] to date over 35 kinase inhibitors have gained regulatory approval from the U.S. Food and Drug Administration.[4–6]

The general strategy in kinase inhibitor drug discovery is focused on targeting the ATP-binding pockets of the catalytic domain. Most kinase inhibitors are reversible ATP-competitive inhibitors, which can either bind to the active form of the kinase ("type I" inhibitors), the inactive form ("type II" inhibitors), or an allosteric binding site outside the ATP-binding pocket of the kinase enzyme ("type III & IV" inhibitors). Reversible kinase inhibitors bind to their preferred targets of interest through non-covalent interactions like hydrogen bonding and dispersion interactions. These inhibitors face issues with potency and duration of therapeutic action due to transient inhibition mechanisms and competition with high intracellular ATP concentrations. This can result in decreased *in vivo* pharmacological activity. Additionally, the structure of the ATP-binding pocket is largely conserved across the kinase family, so some reversible kinase inhibitors may also bind to kinase proteins other than their intended targets. This raises the potential for off-target inhibition. For example, the multi-target reversible kinase inhibitors sunitinib (*Sutent*) and sorafenib (*Nexavar*) have a wide range of toxic side effects owing to promiscuity in targeting multiple kinase enzymes.[7]

Recently, there has been renewed interest among medicinal chemists in developing kinase inhibitors that bind covalently to their targets.[8–18] Targeted covalent inhibitors (TCIs) augment conventional non-covalent protein–ligand interactions with a covalent linkage with a side-chain of the target.[8,10,16] As a result, TCIs can achieve better target efficacy, a longer residence time, and a prolonged duration of therapeutic action in comparison to their conventional non-covalent counterparts. Additionally, TCIs can have exceptionally high target selectivity because in order to achieve covalent inhibition of a desired target, a reactive amino acid must be in a favorable position where a covalent linkage can be formed with the electrophilic moiety of the bound ligand. This is particularly advantageous in the kinase family, where there is a high active-site homology but the targeted cysteine residues are poorly conserved. As a result, the number of proteins in this family that share a cysteine residue are relatively small, limiting the risk of off-target inhibition.[17]

The mechanism of covalent kinase inhibition involves the reaction of an electrophilic warhead of the ligand with a nucleophilic amino acid side-chain in the active site of the targeted kinase. Non-catalytic cysteine residues in or near the active sites of protein kinases have been the primary targets of this approach.[19,20] Cysteine residues possess a chemically-reactive thiol (–SH) moiety that can react with a diverse array

of electrophiles.[21] The reactivity of a cysteine thiol group is complex, with steric, hydrophobic, and electronic factors; however, the propensity for the thiol to be deprotonated is generally the prime determinant of its reactivity towards electrophiles.[22] One successful group of TCI's feature an electrophilic acrylamide warhead, which undergoes a thio-Michael addition with the target cysteine. In the mechanism of this addition, which is illustrated in Figure 4.1, the cysteine thiol is deprotonated to form a thiolate before it reacts with an acrylamide functional group of the inhibitor.[23] Because of the deprotonation step, the rate of covalent modification is dependent on the pK_a of the cysteine residue; more reactive cysteines tend to have lower pK_a's due to the increased probability of the cysteine being in the nucleophilic thiolate state. Accordingly, cysteines with low pK_a's in principle are more susceptible to covalent modification. Thus, calculating the pK_a's of cysteines in kinases could help identify druggable cysteine targets that are prone to covalent modification, ultimately leading to the design of more potent and selective kinase inhibitors for therapeutic intervention.

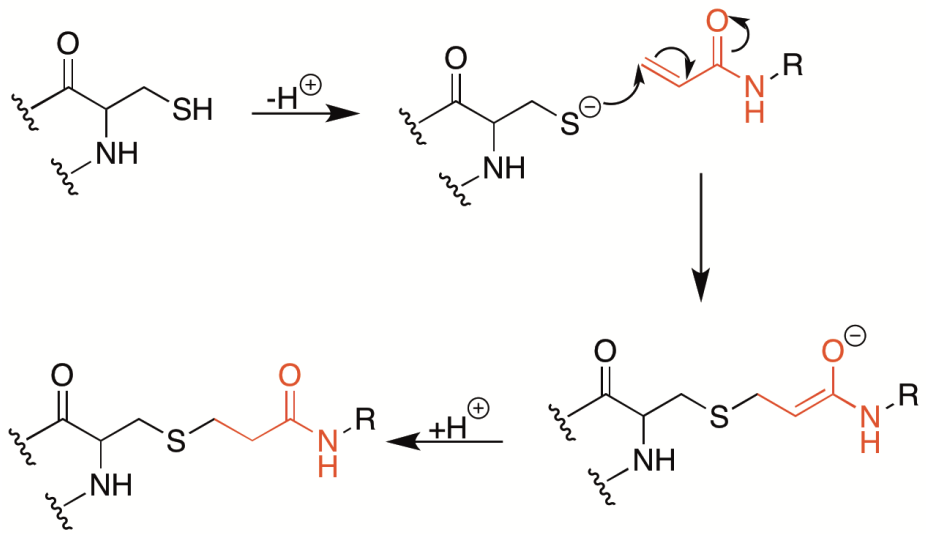

Figure 4.1: Mechanism of covalent modification of cysteine thiol side-chain by an acrylamide warhead moiety (Shown in Red)

Several structure-guided bioinformatics studies have identified cysteines within the human kinome that could be targeted by covalent-modifier drugs.[19,24–27] In a landmark study, Taunton and coworkers[24] employed a structural bioinformatics approach to design selective covalent inhibitors of ribosomal protein S6 kinases (RSKs), taking

advantage of two distinct modes of selectivity: a threonine gatekeeper and a non-conserved solvent-exposed cysteine in the glycine-rich loop region of the ATP-binding pocket. The designed inhibitors were potent and selective for RSK1 and RSK2 kinases over other closely related structural kinase paralogs.[24] Subsequently, Gray and coworkers identified approximately 200 kinases within the human kinome that have a cysteine located in or near the ATP-binding pocket, which could be targeted by a covalent modifier.[25,28] More recently, predictive algorithms, statistical analysis, and computationally-derived properties have been employed to identify which cysteine residues in kinases are optimal targets for a TCI.[26,29–31] A recent study by Bourne and coworkers[30] combined a function-site interaction fingerprint method and density functional theory calculations to explore covalently-modifiable cysteines and their positions across the section of the human kinome where crystal structures are available. The positions of accessible covalent-modifiable cysteines were classified into five regions, namely: phosphate-binding loop (P-loop), roof of binding pocket, front pocket, catalytic loop, and Asp-Phe-Gly (DFG) motif (Figure 4.2).[30] We have used this classification system in our study.

While these structural and bioinformatics studies have identified cysteine residues that have appropriate locations for forming a covalent bond with an inhibitor, no study to date has rigorously assessed if these cysteines will react with an inhibitor at a rate that is sufficiently fast for the covalent linkage to be formed on the necessary timescale. Knowledge of the pK_a's of targetable cysteines in kinases would help to define the intrinsic reactivity of a given cysteine towards an electrophilic inhibitor and provide one of the parameters of the pharmacokinetics of these drugs. These data will guide medicinal chemists in designing drugs that specifically inhibit reactive cysteine targets and avoid wasting effort targeting cysteines that are not reactive.

In this study, we report the calculated pK_a's of 29 druggable cysteine targets in kinases that are active or potential targets for covalent inhibition. Important oncogenic mutants are also included. The pK_a's of the cysteines were calculated using advanced molecular dynamics-based methods, namely, thermodynamic integration[32] and constant-pH molecular dynamics[33,34] simulations in explicit solvent. The kinase structures investigated span the family of the human kinome. Cysteines residues within these kinase structures were chosen based on previous structural analysis studies or where a covalent inhibitor has been reported and confirmed experimentally.

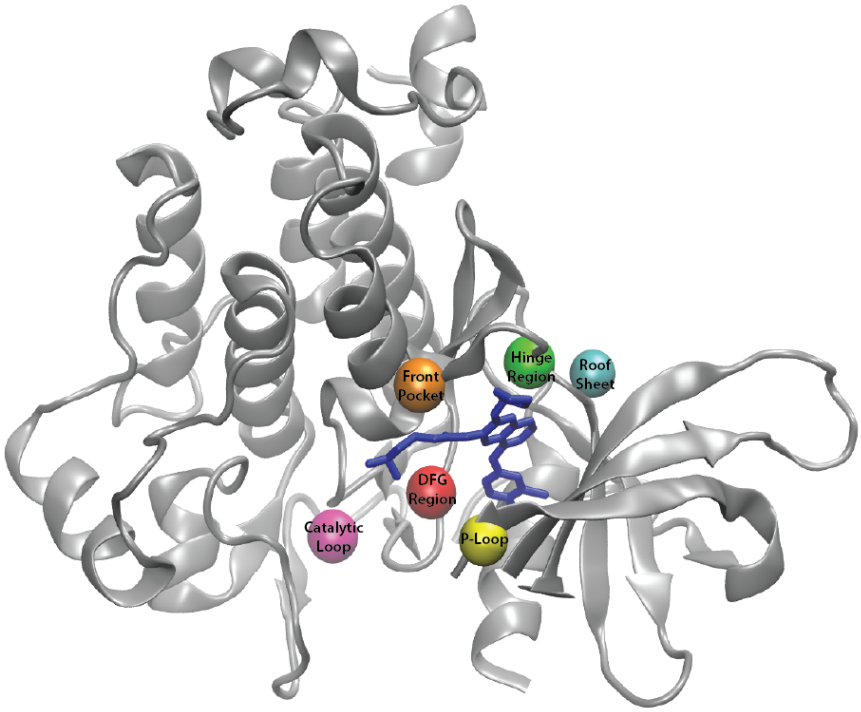

Figure 4.2: The kinase domain of EGFR (PDB ID: 4G5J) in complex with covalent-modifier drug, afatinib (blue). The locations of the target cysteine residues can be categorized into the catalytic loop (magenta), DFG region (red), front pocket (orange), P-loop (yellow), hinge region (green), and the roof sheet (cyan).

4.3 Theory and Methods

Protein Kinases. The initial coordinates of the protein kinase structures used in the molecular dynamics (MD) simulations were collected from the protein data bank (PDB).[35] Atomic coordinates missing from these structure were assigned using SWISS-MODEL homology modelling.[36] Druggable cysteines within these structures were identified by searching the literature for reports of the covalent modification of kinase proteins. In most cases, covalent modification of cysteine targets has been confirmed experimentally, either by mass spectrometry or X-ray crystallographic data. The kinase proteins investigated in this study and targeted cysteine residues are collected in Table 4.1. The majority of the druggable cysteines considered were primarily positioned in three kinase segments, namely: front pocket, P-loop, and DFG region.

PDB identifiers and details of kinase structure models are included in the supplementary information of Ref. 37, (Table S3).

4.3.1 Replica-Exchange Thermodynamic Integration (RETI)

The method of Thermodynamic integration[32] (TI) provides a means of calculating the relative free energy difference between the protonated and deprotonated forms of a target residue in the protein. Replica-exchange molecular dynamics[38–40] improves the convergence and accuracy of the free energies calculated using TI.[41,42]

The pK_a's of the kinase cysteines were calculated using all-atom RETI method in explicit solvent with the CHARMM36 force field.[43] In the RETI approach, the relative Gibbs energies of the thiol and thiolate states of the cysteine residues are calculated in the protein kinase environment and in a reference model system, (Eqn. 4.2). Both acid dissociation reactions are performed in aqueous solution. The reference model is a blocked alanine pentapeptide that contains a cysteine residue: i.e., Ac-$(Ala)_2$-Cys-$(Ala)_2$-NH_2. The end groups of the reference model compound were capped with acetyl and amide functionalities to avoid artifacts from charged termini. Cysteine pK_a's were determined by calculating the shift in pK_a from the reference pK_a value, which is experimentally known. This can be formally expressed as:

$$\mathrm{pK_a(Cys)} = \mathrm{pK_a^{ref}} + \Delta\mathrm{pK_a} \tag{4.1}$$

$$\Delta\mathrm{pK_a} = \frac{1}{2.303 k_B T}(\Delta G_{protein} - \Delta G_{model}) \tag{4.2}$$

where pK_a^{ref} refers to the known reference pK_a value of the cysteine residue in aqueous solution, ($pK_a^{ref} = 8.55 \pm 0.03$).[44] ΔpK_a results from the difference in the intermolecular interactions experienced by the thiol and thiolate states of the cysteine in the protein environment and in the model system.

Computational prediction of the pK_a's of ionizable residues in proteins remains a challenging problem and it is particularly difficult to predict cysteine pK_a's.[45] For a test set of 18 cysteine residues in 12 proteins, the RETI explicit solvent approach with the CHARMM36 force field outperformed all the other methods that were evaluated, yielding an RMSD error of 2.4 pK units.[45] Hybrid quantum mechanics/molecular

Table 4.1: Protein Kinases Studied and their Targeted Cysteine Positions

Protein Kinase	**Cys Residue**	**Region**
BMX	496*	front pocket
BRAF	532	hinge region
BRAF (V600E)	532	hinge region
BTK	481*	front region
c-KIT	788*	catalytic loop
c-Src	277*	P-loop
EGFR	797*	front pocket
EGFR (L858R)	797*	front pocket
EGFR (T790M)	797*	front pocket
EGFR (T790M/L858R)	797*	front pocket
ERK2	166*	DFG region
FGFR1	488*	P-loop
FGFR4	477*	P-loop
FGFR4	552*	hinge region
FLT3	828*	DFG region
GSK3β	199*	DFG region
HER2/ErbB2	805*	front pocket
HER3/ErbB3	721	roof sheet
HER4/ErbB4	778*	front pocket
IKKβ	46*	catalytic loop
IKKβ	179*	activation loop
ITK	442*	front pocket
JAK3	909*	front pocket
JNK2	116*	front pocket
JNK3	154*	front pocket
MEK1/MAP2K1	207*	DFG region
MSK1	440	P-loop
NEK2	22*	P-loop
PDGFRα	814*	catalytic loop
RSK1	432*	P-loop
RSK2	436*	P-loop
TAK1	174*	DFG region
VEGFR2	1045*	DFG refgion

* Cysteine residue confirmed to be covalently modified by experimental X-ray crystallographic or mass spectrometry experiments. Literature reports used to inform covalent modification of cysteines have been included in supplementary Table S3 of Ref. [37].

mechanics (QM/MM) MD simulations have shown that the CHARMM36 force field accurately predicts the experimental hydration structure of a model cysteine thiolate.[46] Benchmark of the RETI method specifically for kinase cysteine pK_a prediction show a similar deviation from experiment (see Supplementary Table S1 in Ref. 37) The method predicted the correct direction in pK shifts for the active site cysteine (Cys283) in wild-type and mutant variants of creatine kinase, with an overall root-mean-square deviation of 1.4 pK units. We note that the pK_a's reported in this study are not expected to be quantitatively accurate, but rather are suggestive of general trends.

Following the RETI pK_a calculation protocol employed in our cysteine pK prediction benchmark study,[46] the GROMACS 5.1.4 molecular dynamics software package[47] was used to perform all the RETI simulations on the protein kinases studied in this work. The CHARMM36[43] all-atom protein force field was used to generate structural models of both the reference model and the kinase system. Titratable residues were assigned protonation states corresponding to their default values at neutral pH. The initial protein kinase structure was centered and solvated in a cubic periodic box, with a cutoff distance of 10 Å from the edge of the box. The simulation cell was neutralized with Na^+ and Cl^- ions at a concentration of approximately 0.10 M. TIP3P water model[48] was used for all the simulations and the system was kept at a temperature of 298.15 K and pressure of 100 kPa using the velocity rescaling thermostat[49] and the Parrinello-Rahman barostat.[50,51] The LINear Constraint Solver algorithm[52] (LINCS) was used to constrain covalent bonds involving hydrogen. Long-range electrostatic interactions were treated using the Particle Mesh Ewald (PME) method.[53,54] A grid spacing of 1.0 Å was used for all simulation cells.

After system preparation, the kinase structure was subjected to the steepest descent energy minimization to eliminate any steric clashes or structural irregularities within the model system. Following that, a 20 ns equilibration run was performed in the canonical ensemble (NVT) followed by an equilibration in the isothermal-isobaric ensemble (NpT), in that order. A time step of 2 fs was used for all the simulations, with a reference temperature and pressure of 298.15 K and 100 kPa, respectively. After system equilibration in the NVT and NpT ensemble, replica exchange thermodynamic integration runs were performed to calculate the relative free energies of the thiol and thiolate states of the selected cysteine. Each free energy calculation is

comprised of 11 windows with λ values ranging from 0.0–1.0, in increments of 0.1. Exchanges between λ states were attempted between neighboring replicas every 1.0 ps. The average exchange probability between replicas was in the 0.10–0.20 range. Gibbs energies were calculated from the RETI data using the weighted histogram analysis method, g_wham.[55] All replicas were run for 12 ns, with the first 2 ns discarded as equilibration. For each kinase model system, the simulations were performed in triplicates. The computed pK_a's reported are the averages of the three independent replicates.

4.3.2 Constant-pH Molecular Dynamics (CpHMD)

CpHMD methods[33,34,56–64] are capable of simulating pH-induced conformational changes while naturally accounting for the variation in protonation states of titratable residues. Additionally, the influence of nearby ionizable residues and coupled protonation states on the pK_a's of titratable residues can be correctly captured by these simulations. In our RETI-pK_a approach, the protonation state of all residues other than the target cysteines of interest are fixed. As a result, contributions of other protonation states on the pK_a's of targeted cysteines are not captured by these simulations. These effects are particularly significant for ionizable residues with coupled pK_a's. We adopted CpHMD methods to investigate the effect of cooperativity between cysteines and their nearby ionizable residues. Of particular interest for study were the structural models of protein kinases EGFR and JAK3. Targetable cysteines within these kinase models (i.e., Cys797 in EGFR and Cys909 in JAK3) have nearby ionizable residues like Asp, whose influence on cysteine pK_a's might not be fully captured by our RETI-pK_a calculations. Supplementary Table S5 in Ref. 37 lists charged ionizable residues that are within 7.0 Å from target cysteines in the set of protein kinases studied.

The CpHMD simulations were performed using two distinct programs; namely, Amber's[65] implementation of replica exchange in the pH-dimension[33] (pH-REMD) and the hybrid nonequilibrium Molecular Dynamics/Monte Carlo (neMD/MC) constant pH approach[34] implemented in the MD program NAMD.[66] The same structures used in the RETI-pK_a calculations were used for the pH-REMD and hybrid neMD/MC calculations. In the pH-REMD approach, simulations of multiple independent replicas are performed at different solution pH values but at the same temperature, and attempts are made to exchange pHs between replicas. The neMD/MC

approach on the other hand, consists of carrying out short nonequilibrium molecular dynamics switching trajectories to generate physically plausible configurations with changed protonation states that are subsequently accepted or rejected according to a Metropolis Monte Carlo criterion.[56]

Constant pH-REMD

The constant pH-REMD simulations were performed in explicit solvent, following the protocol developed by Roitberg and coworkers.[33] Protein kinase models were parametrized using the Amber ff99SB[67] protein force field. Some titratable residues of interest in protein kinase models (i.e., Asp, Glu, His) were modified to properly match titratable residue names used by the program. The tleap program of AmberTools 16 program suite was used to prepare the necessary topology and coordinate files. The intrinsic solvent radius of titratable Asp and Glu residues were reduced by 0.2 Å to compensate for the effect of having 2 dummy protons present on each of the carboxylate oxygen in the syn- and anti-positions.[60] The protein was solvated in a truncated octahedron water box with 10 Å buffer of TIP3P-model water surrounding the protein in each direction. All simulations were performed using either the pmemd or sander module of Amber 14/16 program.[68] The simulations were performed in triplicate and the seed for the random number generator was set from the computer clock to avoid synchronization artifacts.

The simulations followed three initial standard steps before the pH-REMD simulations were performed. The three standard steps include: energy minimization, heating, and equilibration or system relaxation. Langevin dynamics and Berendsen barostat were used to maintain a constant temperature and pressure of the simulation cells. NaCl salt concentration of 0.10 M was maintained using the igb=2 GB model,[69] which has been parameterized for explicit solvent simulations. The SHAKE algorithm[70][71] was used to constrain hydrogen bonds. The Particle Mesh Ewald method was used for treating long-range electrostatics and a 8 Å cutoff was used in the calculation of Lennard-Jones interactions. The minimization was run for 5000 cycles. Following that, the system was heated at a constant volume, varying the target temperature linearly from 10 K to 300 K over 400 ps. The simulation system structure was then equilibrated for 10 ns molecular dynamics simulations in the isothermal-isobaric ensemble (NpT). This equilibrated structure was used as the starting point

for the constant pH-REMD simulations.

The pH-REMD simulations typically consisted of 16 replicas spanning the pH range of interest, in increments of 1.0 pH unit. In the case of EGFR kinase, the replicas were constructed to span a pH range of 0.0–15.0. Replica exchange attempts between adjacent replicas were made every 2500 steps (5 ps) and the protonation state changes were attempted using 100 step (200 fs) non-equilibrium trajectories for all pH-REMD simulations. The simulation was run for a total of 10 ns per replica. The results from the simulation were analyzed using the cphstats program from the Amber program suite. The deprotonated fraction (f_{deprot}) and pH for each individual replica were fitted to the Hill equation, (Eqn. 4.3). The Marquardt-Levenberg fitting equation was then used to calculate the pK_a and Hill coefficient (n), which was then plotted to derive titration curves for ionizable residues of interest. The pH-REMD simulations were run in triplicate and the final pK_a reported is the average of the three independent pK runs performed.

$$f_{deprot} = \frac{1}{1 + 10^{n(pK_a - pH)}} \tag{4.3}$$

Hybrid neMD/MC

The neMD/MC constant pH simulations were performed in explicit solvent using the CHARMM36 protein force field, implemented in NAMD 2.12 program.[66] Kinase model systems were solvated in a water box using the VMD solvate package,[72] with water box dimensions having a 10 Å layer of thickness between the box boundaries and the minimum and maximum coordinates of the protein. Na^+ and Cl^- ions were introduced into the system to neutralize excess charge and the concentration was set to 0.10 M. Periodic boundary conditions were employed using PME electrostatics method and Lennard-Jones interactions were smoothly truncated at 12 Å, using a switching function from 10–12 Å. The SHAKE algorithm was used to constrain covalent bonds containing hydrogen atoms and the simulation time step was 2 fs. A Langevin thermostat of 298.15 K with a damping coefficient of 1 ps^{-1} was used for equilibration.

For each simulation, the system was first minimized (1000 steps), followed by 20 ns of equilibrium molecular dynamics. The equilibration sampled an isothermal-isobaric

ensemble (NpT) with a temperature of 298.15 K and a pressure of 101.325 kPa using Langevin dynamics. The equilibrated model system was used as the starting point for the constant pH molecular dynamics simulations. The theoretical background of the neMD/MC constant-pH MD approach employed in this study is described in detail in a publication by Roux and coworkers.[34] Given the large number of titratable residues present in the protein kinase models, residues of particular interest (Cys and neighboring Asp) were assigned a proposal weight of 10.0, while the remaining titratable residues were assigned the default proposal weight of 1.0 during simulations. Inherent pK_a values for all titratable residues were assigned using their reference pK_a values in bulk solution, as calibrated for by the method.[34] All simulations attempted protonation state moves every 10 ps with switching times of 15 ps. All simulations were repeated three times and were run for 20 ns within pH values where ionizable residues of interest titrate. Titration curves and error estimates from simulation outputs are computed using a Python-based utility, cphanalyze, available through the PyNAMD package. The PyNAMD library is accessible via the URL: `https://github.com/radakb/pynamd`. The computed titration curves were fitted to the Hill equation, Eqn. 4.3. The reported pK_a values are the averages of three independent replicates.

4.4 Results and Discussion

The pK_a's of the targetable cysteines using the RETI method range from values as low as 7 pK units to as high as 24 pK units (Figure 4.3). Most of the cysteine residues have pK_a's that are elevated above the pK_a of a free cysteine thiol in solution, 8.6.[44] Only the targeted cysteine residues of c-Src, FGFR4, and JNK2 yielded predicted pK_a's lower than the solution cysteine pK_a. In principle, cysteine residues with low pK_a's are more susceptible to covalent modification because the thiol side-chain is easily deprotonated and is more likely to exist in the reactive nucleophilic thiolate state. There was no systematic trend between predicted cysteine pK_a's and their positions in the kinase segment, although cysteines in the catalytic loop of protein kinases c-KIT and PDGFRα, reported the highest pK_a's (i.e. > 20). In reality, these proteins could likely undergo a conformational change such that the deprotonation could occur at a lower pH or covalent modification could occur through a non-ionic mechanism.

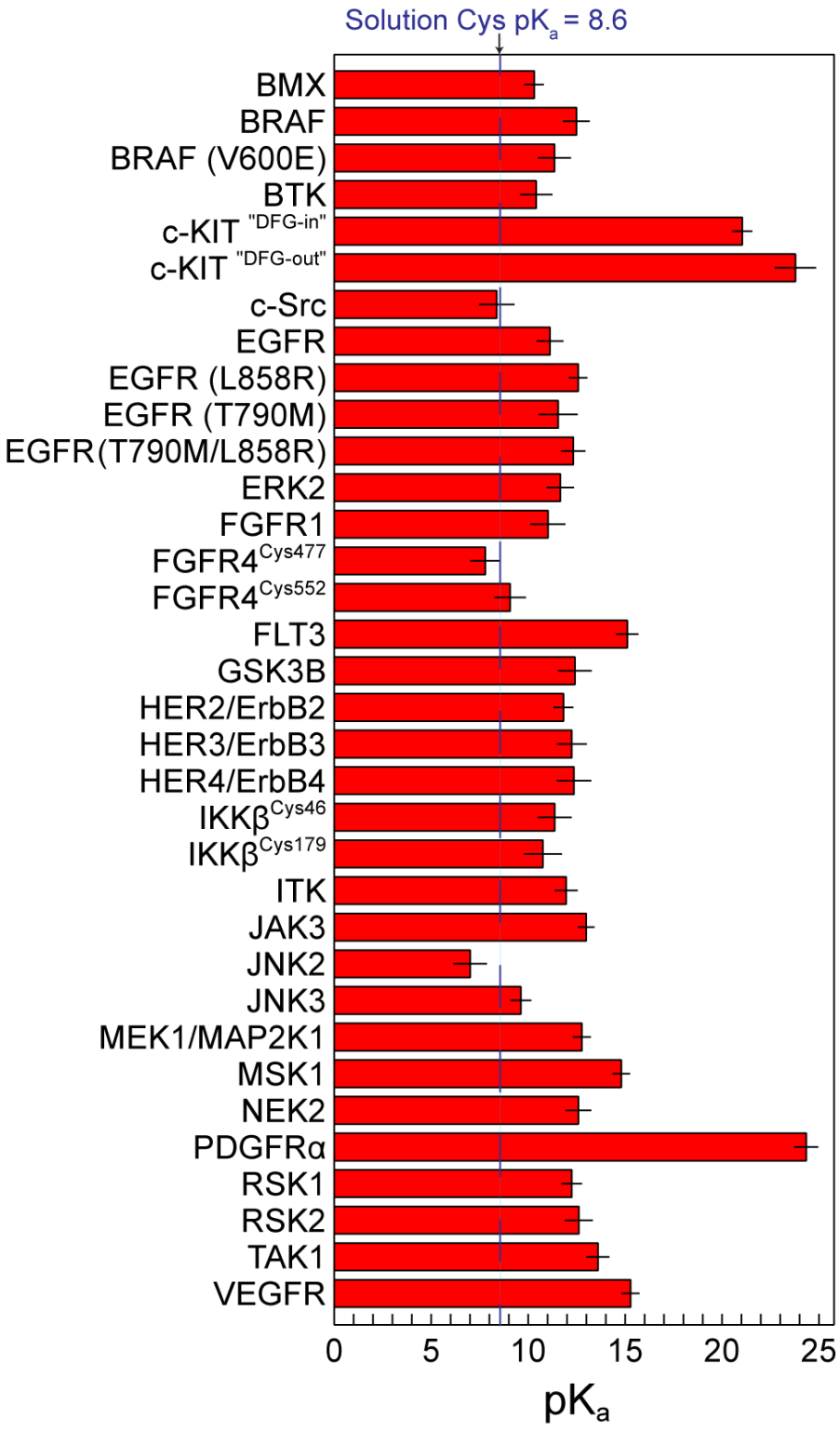

Figure 4.3: Calculated pK$_a$'s of covalent-modifiable cysteines in selected protein kinases using RETI method.

Two particularly important conformational states of protein kinases are the active ("DFG-in") and inactive ("DFG-out") states. These are relevant in c-KIT and PDGFRα, where there has been efforts to develop inhibitors for the DFG-out state.[73,74] In the active conformation, the Asp residue at the N-terminus of the activation loop within the conserved DFG motif points towards the ATP-binding site ("DFG-in"); while in the inactive conformation, this position is occupied by the Phe residue of the DFG motif ("DFG-out"). These local structural differences in kinase DFG motif conformation can impact the physicochemical environment of ionizable residues in functional domains—perturbing their pK_a. To investigate the effect of different kinase conformations on the predicted pK_a's of targetable residues, we calculated the pK_a of Cys788 of c-KIT in the inactive "DFG-out" (PDB ID: 3G0E) and active "DFG-in" (PDB ID: 1PKG) conformations; and we find that the "DFG-in" conformation yields a slightly lower pK_a (21.0 ± 0.5) than the "DFG-out" conformation—although still elevated (Figure 4.3). The dislocation of the juxtamembrane domain from its autoinhibitory position into solvent in the active "DFG-in" conformation is a likely cause for the difference in Cys788 pK_a's for both "DFG-in" and "DFG-out" c-KIT kinase conformations.[75] This indicates that different kinase conformation can have an effect on the pK_a's of cysteine residues near the catalytic loop region, especially if it involves a conformational transition of the DFG motif.[76] As a result, the predicted pK_a's will vary depending on the conformation present in the PDB structure. The predicted pK_a's we report for the targeted cysteine residues in c-KIT and PDGFRα kinases are for their autoinhibited "DFG-out" forms, which bind inhibitors imatinib and sunitinib.[73] The high predicted pK_a's for these kinase cysteines are largely notional and indicative of the physicochemical environment around the cysteines, which appear to be desolvated and buried away in the protein interior. The wide range observed in the acidities of kinase cysteines indicates that the reactivity of targeted cysteine residues can vary greatly across the protein kinase family.

The variation in the cysteine pK_a's is due to differences in the environment around the residue. Thiols have limited intermolecular interactions, but the stability of the thiolate state is sensitive to its interactions with water molecules and other ionizable residues in the protein.[77,78] We computed cysteine thiolate hydration numbers from the trajectories of free energy calculations of the protein kinase models (see Supplementary Table S4 in Ref. 37). A solvent-exposed cysteine thiolate is predicted to have a hydration number of 4.4. The hydration of the cysteine thiolate has a dramatic effect

on the pK_a (Figure 4.4).

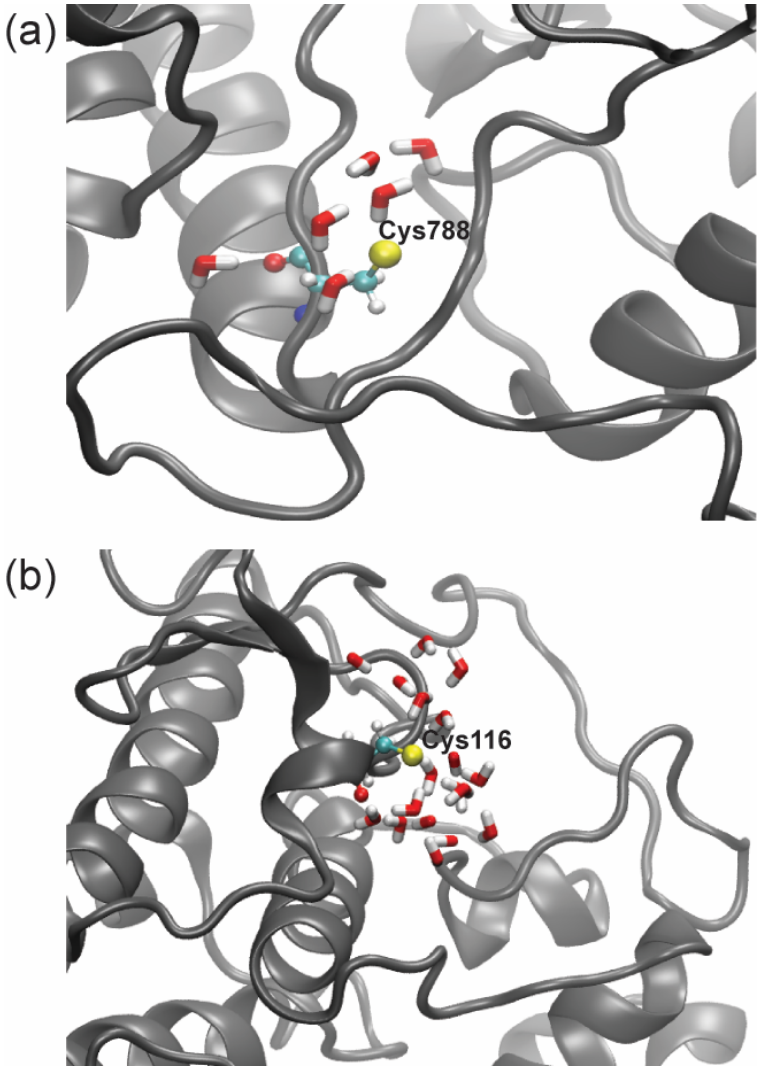

Figure 4.4: Factors perturbing the pK_a of druggable cysteines in protein kinases. (a) Poor solvation of thiolate state of Cys788 in c-KIT kinase (PDB ID: 3G0E) results in an elevated pK_a (23.8). (b) Solvent-exposed Cys116 in JNK2 (PDB ID: 3E7O) kinase reports a predicted pK_a of 7.0. Only water molecules within 5 Å from the target cysteine are shown. The hydration numbers calculated are 3.6 and 5.5 for c-KIT and JNK2 kinase, respectively.

Buried cysteine residues that are poorly solvated have higher pK_a values, while solvent-exposed cysteine residues have lower pK_a values. For example, Cys277 and Cys116 in c-Src and JNK2 kinases are predicted to have relatively low pK_a's and have high hydration numbers of 6.0 and 5.5, respectively (Table 4.2). Conversely, cysteines with low thiolate hydration numbers (i.e., 3–5) typically have pK_a's greater than 11.

Table 4.2: Thiolate Hydration Numbers of Select Kinase Cysteines

Protein kinase	**Cys Residue**	**RETI-pK_a**	**Hydration number**
c-KIT$^{\text{"DFG-out"}}$	788	23.8 ± 1.0	3.6
c-Src	277	8.3 ± 0.9	6.0
EGFR	797	11.1 ± 0.7	4.3
ERK2	166	11.7 ± 0.7	4.0
FGFR4	477	7.8 ± 0.7	4.4
HER2	805	11.8 ± 0.5	3.8
JAK3	909	13.0 ± 0.4	5.4
JNK2	116	7.0 ± 0.8	5.5
JAK3	207	12.8 ± 0.4	4.0
PDGFRα	814	24.3 ± 0.6	3.7

This effect is particularly strong for Cys788 of c-KIT (Figure 4.4 (a)) and Cys814 of PDGFRα, which have very low thiolate hydration numbers and extremely high pK_a's. Notable exceptions to this trend include Cys477 in FGFR4 and Cys909 in JAK3. Cys477 in FGFR4 is predicted to have modest hydration number of 4.4 but a pK_a of only 7.8. Lys644 is near to Cys477, which results in a stabilizing cation-anion interaction with the cysteine thiolate. This results in a relatively low pK_a for Cys477, despite being poorly solvated. On the other hand, Cys909 in JAK3 kinase is predicted to have a high hydration number of 5.4, but an elevated pK_a of 13. Asp912 is close to Cys909 and destabilizes the cysteine thiolate anion, elevating the pK_a.

Inter-residue electrostatic interactions can also significantly shift the pK_a of an amino acid in a protein. Several of the target cysteine residues are in close proximity to an amino acid with an anionic side-chain. This generally results in an increase in the pK_a of the cysteine residue due to electrostatic repulsion between thiolate and the anionic side-chain. For instance, Cys207 of MEK1 kinase neighbors Asp208 and its predicted pK_a is elevated to 12.8 (Figure 4.5(a)). In ERK2 kinase, Asp167 neighbors Cys166 and this yields an elevated pK_a of 11.7 for Cys166 (Figure 4.5(b)). Both thiolates in the above protein kinases have similar hydration numbers (Table 4.2), but the influence and relative proximity of inter-residue electrostatic interactions within the cysteine microenvironment is a contributing factor to their elevated pK_a's.

The predicted shifts in cysteine pK_a's versus the thiolate hydration numbers from the RETI calculations are plotted in Figure 4.6. Targetable cysteines surrounded by negatively charged amino acid side-chains within 6.5 Å are represented as red

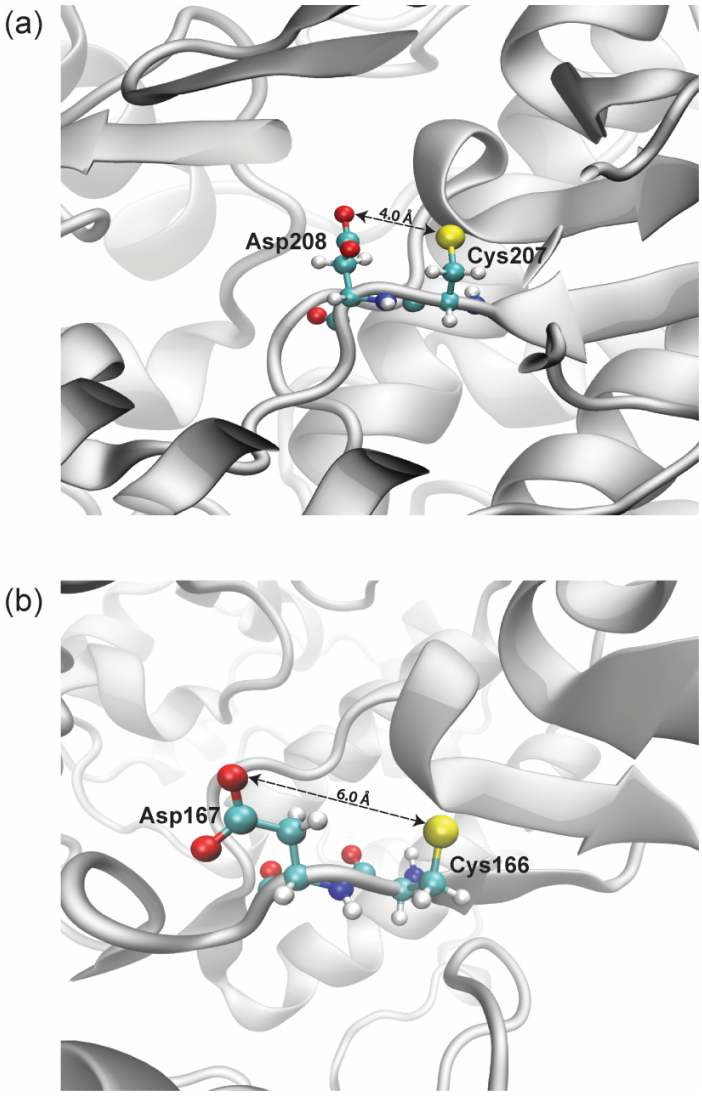

Figure 4.5: Electrostatic effects of neighboring Asp on the pK_a's of select kinase cysteines. (a) The proximity of Asp208 to Cys207 contributes to the elevated Cys pK_a(12.8) in MEK1 kinase. (b) The thiol side-chain of Cys166 in ERK2 kinase (pK_a= 11.7) is 6.0 Å from the carboxylate anion of Asp167.

circles, while cysteines surrounded by positively charged side-chain groups within 6.5 Å are represented as blue circles. Cysteines with no acidic or basic side-chain groups within 6.5 Å are represented as black circles. As expected, cysteines with nearby acidic residues on average have slightly larger pK shifts than cysteines surrounded by basic residues. Several residues with the lowest hydration numbers have very elevated pK_a shifts, while residues with high hydration numbers have pK_a shifts that are modestly positive or are even shifted in the negative direction in some cases. Although electrostatic interactions with nearby residues and hydration of the thiolate state can clearly have a significant influence on the computed cysteine pK_a's, the diversity in these values highlights that the pK_a of a residue depends on a complex set of factors.

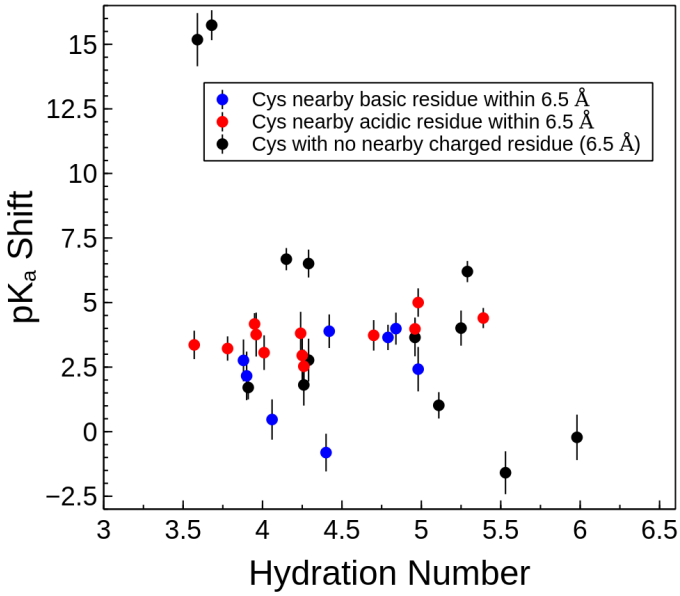

Figure 4.6: pK_a shifts vs. thiolate hydration number for targetable cysteine residues, as predicted by the RETI method. Blue and red circles represent cysteines surrounded by positively-charged and negatively-charged groups within 6.5 Å, respectively. Black circles represent targeted cysteine residues with no nearby charged groups within 6.5 Å.

The observation that the active-site cysteine residues tend to have elevated pK_a's is consistent with other reports that the protonation states of active-site residues can be shifted considerably from their solution values, even when these residues are not catalytic. Isom et al.[79] showed that lysine residues engineered into the active sites of staphylococcal nuclease proteins tended to have low pK_a's. In that case, the poor

hydration of the residues in the active site favored the deprotonated, neutral state of lysine. For cysteine residues, the low solvent exposure favors the neutral thiol form relative to the thiolate form, resulting in an elevated pK_a. This suggests that non-catalytic active-site lysine residues will tend be more reactive towards nucleophiles than amines in solution, while non-catalytic active-site cysteines generally tend to be less reactive than cysteines in solution.[17,80] Paradoxically, the high pK_a's of these cysteine residues may protect them from oxidation, preserving them as targets for reaction with electrophilic drugs.[81]

Recent progress in kinase inhibitor drug discovery has seen the regulatory approval of a number of small-molecule covalent kinase inhibitors.[18,82] Majority of these inhibitors have been successful at targeting noncatalytic cysteine residues near the ATP-binding pocket. Cys797 in EGFR, Cys805 in HER2, and Cys481 in BTK are proven examples where effective, cysteine-targeting covalent inhibitors have been developed successfully.[82] Within the kinases where there is an effective and well-characterized covalent inhibitors, the calculated pK_a's tend to be modestly elevated. BTK, EGFR, and Her2 are inhibited, respectively, by the drugs ibrutinib,[83] afatinib,[84] and neratinib.[85] The targeted cysteine in these kinases are all located in the front pocket region of the active site (Table 4.1), and their calculated pK_a's range between 10 and 11 pK units. The thiolate state of these residues tends to have a moderate hydration number (i.e., 4–5 water molecules). Based on the demonstrated success of drugs that target these residues, front-pocket cysteines appear to have a combination of acidity and positioning that make them well-suited for covalent inhibition. This indicates that a residue with a pK_a of 11 can still be targeted, although complex factors including protein–ligand binding and conformational changes could perturb the pK_a of a targetable residue, making it more amenable to covalent modification.

The pK_a of Cys797 in EGFR is particularly significant for drug development. Several chemotherapy drugs have functioned by inhibiting EGFR through covalent modification of Cys797, including afatinib[84] and osimertinib.[86] Sequence alignment studies have shown that ten other kinases within the human kinome possess a cysteine residue in a structurally similar position as Cys797 of EGFR.[82] The RETI-pK_a method predicts the pK_a of Cys797 in EGFR to be elevated to 11. This is in contrast to the experimental work of Truong et al.,[87] who measured the rate of bromobimane fluorescent labeling of recombinant EGFR over the pH range of 2–6.5 and estimated

the pK_a of Cys797 to be 5.5. This low pK_a value reported was attributed to stabilizing electrostatic interactions between the cysteine thiolate and the backbone dipole formed by the α-helix of residues 799-806.[87] This is a larger effect than has been reported for other cysteine residues at the N-cap position of an α-helix;[88] in comparison, a cysteine engineered at the N-cap position of a 26-residue α-helical segment in myoglobin is only reduced to 6.5.[89] This effect of the α-helix dipole should be partially captured by the molecular mechanical force field used in the RETI-pK_a calculations, although the neglect of induced polarization may cause the cooperative effect of the helix dipole to be underestimated.[90,91]

One of the approximations of the RETI-pK_a calculations is the neglect of cooperativity of side-chain protonation states. For instance, an anionic residue nearby the cysteine would generally raise its pK_a, but if this side-chain was protonated all or part of the time, its effect on the cysteine pK_a would be attenuated or even reversed. In EGFR, it has been hypothesized that Asp800 can act as a general base/general acid in distinct steps of the covalent modification of Cys797, including functioning as a hydrogen bond donor to the thiolate when in its carboxylic acid form.[92,93] To explore the effect of cooperativity between titratable residues, we calculated the titration curves for select cysteines and aspartates in EGFR and JAK3 kinases using two constant pH molecular dynamics (CpHMD) methods; pH-REMD[33] and hybrid neMD/MC.[34] CpHMD takes into account the variation of protonation states in the pK_a calculation—allowing both ionizable residues of interest (i.e., Cys and the nearby Asp) to be titrated simultaneously. These methods allow the protonation state of a residue to change over the course of the simulation. pH-REMD calculates the probability of a transition between two protonation states using an implicit solvent calculation, while the neMD/MC simulation is performed entirely with an explicit solvent representation. Details about the CpHMD methods employed and simulation procedures can be found in the theory and methods section of this chapter. Figure 4.7 shows the representative configuration of the target Cys and nearby Asp residues in both EGFR and JAK3 kinase models.

The percentage of time each residue of interest is deprotonated as a function of pH is plotted in Figures 4.8 and 4.9. The equivalence point of the titration curve of Cys797 in EGFR is evaluated to pK_a value of 13.5 by the pH-REMD method (Figure 4.8 (a)). The hybrid neMD/MC method on the other hand, evaluates the pK_a of Cys797 in EGFR to be 11.5 (Figure 4.8 (b)). Similarly, the equivalence point of the

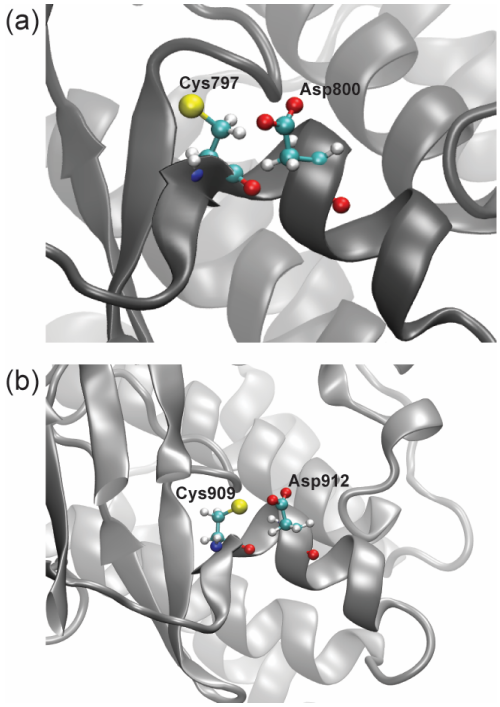

Figure 4.7: Representative configuration of Cys797 and proximal Asp800 in EGFR kinase (a); and Cys909 with nearby Asp912 in JAK3 kinase (b).

titration curve of Cys909 in JAK3, which features a homologous pairing with Asp912, is predicted to be 12.7 by the pH-REMD method (Figure 4.9 (a)) and 11.1 for the neMD/MC method Figure 4.9 (b)).

Although these methods are fundamentally different approaches for calculating ionizable residue pK_a and employ different force field models and parameters for thiol/thiolate cysteines,[46] the computed titration curves agree reasonably well with each other. Furthermore, the calculated pK_a's for Cys797 in EGFR and Cys909 in JAK3 using the constant pH methods are in good agreement with the RETI-pK_a simulation results; which predict pK_avalues of 11.1 and 13.0 for Cys797 and Cys909, respectively. Both the CpHMD and hybrid neMD/MC simulations predict significantly elevated pK_a's for these cysteine residues, with no cooperativity with the nearby Asp residue, so these models are inconsistent with a stable thiolate—aspartic acid pairing.

The titration curves suggest that targetable cysteines present inside the binding sites of these protein kinase models have pK_a's that are higher than in solution. It is also clear from the titration curves that there is no instance where the protonation

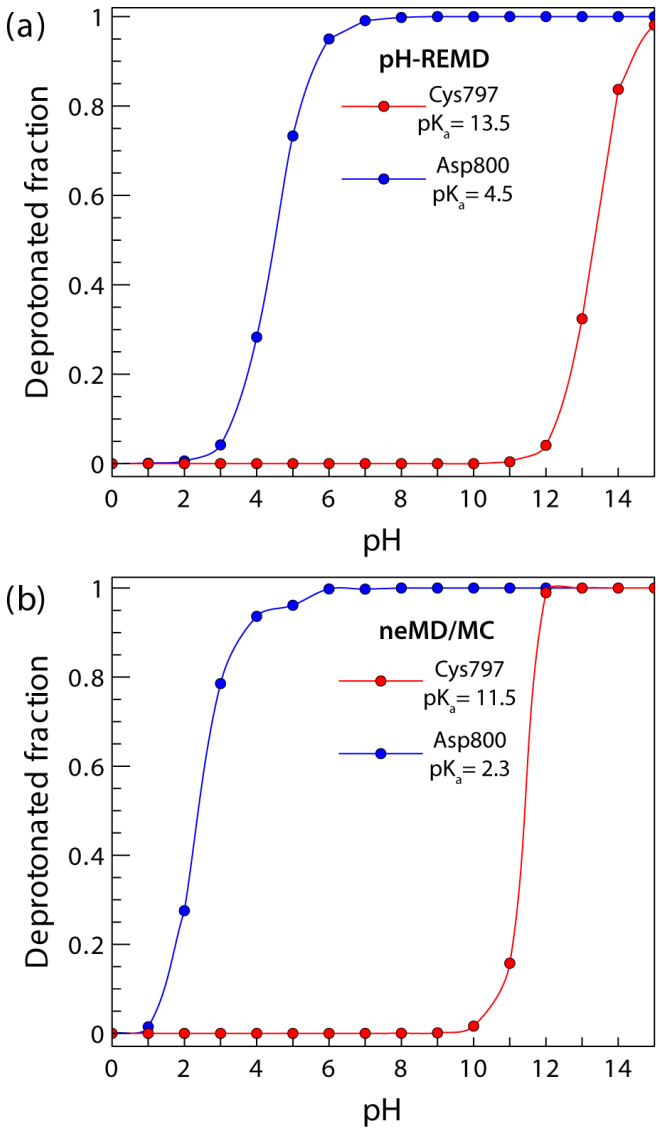

Figure 4.8: Titration curves of Cys797 and Asp800 in wild-type EGFR kinase. The fitted titration curves were generated from deprotonated fractions of pH-REMD (a), and neMD/MC (b), constant pH molecular dynamics simulations. The pK$_a$'s reported are the average of three independent replicates.

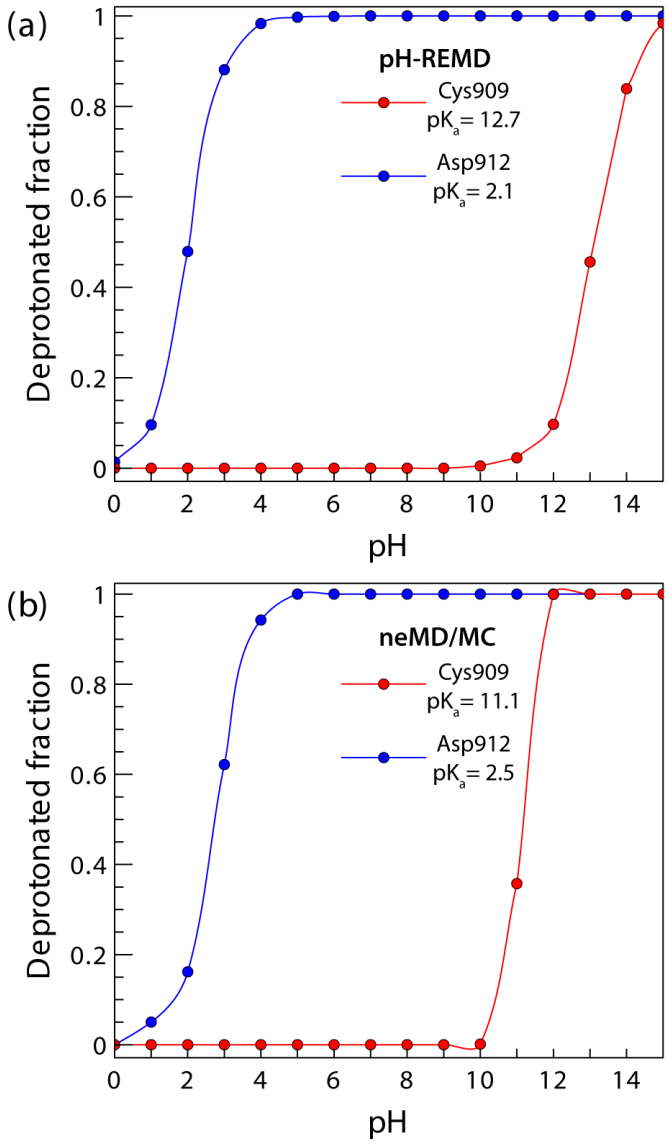

Figure 4.9: Titration curves of Cys909 and Asp912 in JAK3 kinase. The fitted titration curves were generated from deprotonated fractions of pH-REMD (a), and neMD/MC (b), constant pH molecular dynamics simulations. The pK_a's reported are the average of three independent replicates.

states of cysteine and aspartate are coupled or fractionally populated — evident in the characteristic sigmoidal shape pH titration curves observed for both residues (Figures 4.8 and 4.9). This suggests that the transfer of a proton from Cys797 to Asp800 as part of the reaction mechanism of the covalent inhibition of EGFR would be a significantly exergonic step, although the presence of the ligand could affect this.[94] Based on these simulations, we expect that the dominant effect of an acidic residue proximal to a cysteine will be to elevate its pK_a, as observed.

The pK_a of 5.5 for Cys797 in EGFR reported by Truong et al.[87] contradicts the elevated pK_a values calculated by all the molecular dynamics methods employed in this study, including the pK_a's calculated by popular pK prediction web servers: H++[95] (pK_a^{Cys797}=10.5) and PROPKA[96] (pK_a^{Cys797}=10.4). To ensure that the thiolate—carboxylic state was accessed in the molecular dynamics simulations, a simulation was performed where the pK_a of Cys797 was initially assigned the pK_a of 5.5 reported by Truong et al.[87] so that the simulations would begin in this state. After 20-ns of neMD/MC sampling, the protonation state had reverted to the thiol—carboxylate state with an elevated pK_a value of 10 for Cys797, consistent with the results of our earlier CpHMD simulations. This disparity between the reported Cys797 pK_a in EGFR and those calculated by the computational methods merits further investigation.

Given that there is a lack of quantitative experimental data about kinase cysteine residues in general, it would be very challenging to include all possible kinase conformations in the pK_a calculations, although in principle, the protein could adopt different conformations during covalent modification than the crystallographic structures used in our molecular dynamics simulations. Also, limitations of empirical molecular mechanics force fields, protonation state sampling, and convergence issues in computed free energies can introduce errors in the calculated pK_a's. Constant-pH MD simulations that include induced polarization effects and more direct experimental observation of the protonation states of ionizable residues, such as NMR titration, would help unambiguously determine the protonation state of targetable cysteine residues in kinase proteins. This would be particularly valuable for conclusively determining the pK_a of important targets like Cys797 in EGFR and other side-chains to reconcile the apparent inconsistency between computational and experimental results.

4.5 Conclusions

In summary, we have calculated the pK_a's of select kinase cysteine residue targets for covalent-modifier drugs. The pK_a's were computed using rigorous simulation methods with an explicit representation of the solvent. Our analysis suggests that the degree of solvation of the thiolate state and interactions between the thiolate and other amino acids in the binding site are responsible for the perturbation in the pK_a's of druggable cysteines in kinases. The general trend is for the pK_a's of the cysteines to be elevated because they are poorly solvated in comparison to residues on the surface of the protein. The pK_a of a cysteine residue can also be perturbed by electrostatic interactions with other charged residues; nearby anionic residues increase pK_a due to electrostatic repulsion with the cysteine thiolate. These simulations indicate that cysteine residues in kinase active sites will have widely different susceptibility to reactions with an electrophilic drug due to differences in their acidity. Given the trend for active-site kinase cysteines to have elevated pK_a's, the step in the covalent modification mechanism where the thiol is deprotonated will occur at a slower rate than in solution. Computational prediction of the stability of the thiolate state and other intermediate states of covalent modification will enable more rational development of this important class of drugs. The evolution of pK_a prediction methods by improved force fields (e.g., those that include induced polarization), cooperativity of side-chain protonation states, and conformational sampling will allow the acidity of these residues to be predicted more rigorously.

Bibliography

[1] Manning, G. The Protein Kinase Complement of the Human Genome. *Science* **2002**, *298*, 1912–1934.

[2] Lahiry, P.; Torkamani, A.; Schork, N. J.; Hegele, R. a. Kinase mutations in human disease: interpreting genotype-phenotype relationships. *Nat. Rev. Genet.* **2010**, *11*, 60–74.

[3] Cohen, P. Protein kinases — the major drug targets of the twenty-first century? *Nat. Rev. Drug Discovery* **2002**, *1*, 309–315.

[4] Wu, P.; Nielsen, T. E.; Clausen, M. H. FDA-approved small-molecule kinase inhibitors. *Trends Pharmacol. Sci.* **2015**, *36*, 422–439.

[5] Berndt, N.; Karim, R. M.; Schönbrunn, E. Advances of small molecule targeting of kinases. *Curr. Opin. Chem. Biol.* **2017**, *39*, 126–132.

[6] Ferguson, F. M.; Gray, N. S. Kinase inhibitors: the road ahead. *Nat. Rev. Drug Discovery* **2018**, *17*, 353–377.

[7] Kim, A.; Balis, F. M.; Widemann, B. C. Sorafenib and sunitinib. *The oncologist* **2009**, *14*, 800–5.

[8] Potashman, M. H.; Duggan, M. E. Covalent modifiers: An orthogonal approach to drug design. *J. Med. Chem.* **2009**, *52*, 1231–1246.

[9] Singh, J.; Petter, R. C.; Baillie, T. A.; Whitty, A. The resurgence of covalent drugs. *Nat. Rev. Drug Discovery* **2011**, *10*, 307–317.

[10] Lonsdale, R.; Ward, R. A. Structure-based design of targeted covalent inhibitors. *Chem. Soc. Rev.* **2018**, *47*, 3816–3830.

[11] Barf, T.; Kaptein, A. Irreversible Protein Kinase Inhibitors: Balancing the Benefits and Risks. *J. Med. Chem.* **2012**, *55*, 6243–6262.

[12] Kalgutkar, A. S.; Dalvie, D. K. Drug discovery for a new generation of covalent drugs. *Expert Opin. Drug Discovery* **2012**, *7*, 561–581.

[13] Wilson, A. J.; Kerns, J. K.; Callahan, J. F.; Moody, C. J. Keap calm, and carry on covalently. *J. Med. Chem.* **2013**, *56*, 7463–7476.

[14] Sanderson, K. Irreversible kinase inhibitors gain traction. *Nat. Rev. Drug Discovery* **2013**, *12*, 649–651.

[15] Bauer, R. A. Covalent inhibitors in drug discovery: From accidental discoveries to avoided liabilities and designed therapies. *Drug Discov. Today* **2015**, *20*, 1061–1073.

[16] Baillie, T. A. Targeted Covalent Inhibitors for Drug Design. *Angew. Chem. Int. Ed.* **2016**, *55*, 13408–13421.

[17] Lagoutte, R.; Patouret, R.; Winssinger, N. Covalent inhibitors: an opportunity for rational target selectivity. *Curr. Opin. Chem. Biol.* **2017**, *39*, 54–63.

[18] Zhao, Z.; Bourne, P. E. Progress with covalent small-molecule kinase inhibitors. *Drug Discov. Today* **2018**, *23*, 727–735.

[19] Leproult, E.; Barluenga, S.; Moras, D.; Wurtz, J.-M.; Winssinger, N. Cysteine mapping in conformationally distinct kinase nucleotide binding sites: application to the design of selective covalent inhibitors. *J. Med. Chem.* **2011**, *54*, 1347–55.

[20] Serafimova, I. M.; Pufall, M. A.; Krishnan, S.; Duda, K.; Cohen, M. S.; Maglathlin, R. L.; McFarland, J. M.; Miller, R. M.; Frödin, M.; Taunton, J. Reversible targeting of noncatalytic cysteines with chemically tuned electrophiles. *Nat. Chem. Biol.* **2012**, *8*, 471–6.

[21] Shannon, D. A.; Weerapana, E. Covalent protein modification: The current landscape of residue-specific electrophiles. *Curr. Opin. Chem. Biol.* **2015**, *24*, 18–26.

[22] Ferrer-Sueta, G.; Manta, B.; Botti, H.; Radi, R.; Trujillo, M.; Denicola, A. Factors affecting protein thiol reactivity and specificity in peroxide reduction. *Chem. Res. Toxicol.* **2011**, *24*, 434–450.

[23] Smith, J. M.; Rowley, C. N. Automated computational screening of the thiol reactivity of substituted alkenes. *J. Comput.-Aided Mol. Des.* **2015**, *29*, 725–735.

[24] Cohen, M. S.; Zhang, C.; Shokat, K. M.; J., T. Structural Bioinformatics-Based Design of Selective, Irreversible Kinase Inhibitors. *Science* **2005**, *308*, 1318–1321.

[25] Liu, Q.; Sabnis, Y.; Zhao, Z.; Zhang, T.; Buhrlage, S. J.; Jones, L. H.; Gray, N. S. Developing irreversible inhibitors of the protein kinase cysteinome. *Chem. Biol.* **2013**, *20*, 146–159.

[26] Awoonor-Williams, E.; Walsh, A. G.; Rowley, C. N. Modeling covalent-modifier drugs. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* **2017**, *1865*, 1664–1675.

[27] Chaikuad, A.; Koch, P.; Laufer, S. A.; Knapp, S. The Cysteinome of Protein Kinases as a Target in Drug Development. *Angew. Chem. Int. Ed.* **2018**, *57*, 4372–4385.

[28] Zhang, J.; Yang, P. L.; Gray, N. S. Targeting cancer with small molecule kinase inhibitors. *Nat. Rev. Cancer* **2009**, *9*, 28–39.

[29] Zhang, Y.; Zhang, D.; Tian, H.; Jiao, Y.; Shi, Z.; Ran, T.; Liu, H.; Lu, S.; Xu, A.; Qiao, X.; Pan, J.; Yin, L.; Zhou, W.; Lu, T.; Chen, Y. Identification of covalent binding sites targeting cysteines based on computational approaches. *Mol. Pharmaceutics* **2016**, *13*, 3106–3118.

[30] Zhao, Z.; Liu, Q.; Bliven, S.; Xie, L.; Bourne, P. E. Determining Cysteines Available for Covalent Inhibition Across the Human Kinome. *J. Med. Chem.* **2017**, *60*, 2879–2889.

[31] Zhang, W.; Pei, J.; Lai, L. Statistical Analysis and Prediction of Covalent Ligand Targeted Cysteine Residues. *J. Chem. Inf. Model.* **2017**, *57*, 1453–1460.

[32] Straatsma, T. P.; Berendsen, H. J. C. Free energy of ionic hydration: Analysis of a thermodynamic integration technique to evaluate free energy differences by molecular dynamics simulations. *J. Chem. Phys.* **1988**, *89*, 5876.

[33] Swails, J. M.; York, D. M.; Roitberg, A. E. Constant pH replica exchange molecular dynamics in explicit solvent using discrete protonation states: Implementation, testing, and validation. *J. Chem. Theory Comput.* **2014**, *10*, 1341–1352.

[34] Radak, B. K.; Chipot, C.; Suh, D.; Jo, S.; Jiang, W.; Phillips, J. C.; Schulten, K.; Roux, B. Constant-pH Molecular Dynamics Simulations for Large Biomolecular Systems. *J. Chem. Theory Comput.* **2017**, *13*, 5933–5944.

[35] 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*, 235–42.

[36] Arnold, K.; Bordoli, L.; Kopp, J.; Schwede, T. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics* **2006**, *22*, 195–201.

[37] Awoonor-Williams, E.; Rowley, C. N. How Reactive are Druggable Cysteines in Protein Kinases? *J. Chem. Inf. Model.* **2018**, *58*, 1935–1946.

[38] Sugita, Y.; Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **1999**, *314*, 141–151.

[39] Sugita, Y.; Kitao, A.; Okamoto, Y. Multidimensional replica-exchange method for free-energy calculations. *J. Chem. Phys.* **2000**, *113*, 6042–6051.

[40] Fukunishi, H.; Watanabe, O.; Takada, S. On the Hamiltonian replica exchange method for efficient sampling of biomolecular systems: Application to protein structure prediction. *J. Chem. Phys.* **2002**, *116*, 9058.

[41] Woods, C. J.; Essex, J. W.; King, M. a. Enhanced configurational sampling in binding free-energy calculations. **2003**, 13711–13718.

[42] Woods, C. J.; Essex, J. W.; King, M. A. The Development of Replica-Exchange-Based Free-Energy Methods. *J. Phys. Chem. B* **2003**, *107*, 13703–13710.

[43] Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E. M.; Mittal, J.; Feig, M.; MacKerell, A. D. Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone , ψ and Side-Chain χ 1 and χ 2 Dihedral Angles. *J. Chem. Theory Comput.* **2012**, *8*, 3257–3273.

[44] Thurlkill, R. L.; Grimsley, G. R.; Scholtz, J. M.; Pace, C. N. pK values of the ionizable groups of proteins. *Protein Sci.* **2006**, *15*, 1214–1218.

[45] Awoonor-Williams, E.; Rowley, C. N. Evaluation of Methods for the Calculation of the pKa of Cysteine Residues in Proteins. *J. Chem. Theory Comput.* **2016**, *12*, 4662–4673.

[46] Awoonor-Williams, E.; Rowley, C. N. The hydration structure of methylthiolate from QM/MM molecular dynamics. *J. Chem. Phys.* **2018**, *149*, 045103.

[47] Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2015**, *1-2*, 19–25.

[48] Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926.

[49] Bussi, G.; Donadio, D.; Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **2007**, *126*, 014101.

[50] Parrinello, M.; Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **1981**, *52*, 7182–7190.

[51] Nosé, S.; Klein, M. Constant pressure molecular dynamics for molecular systems. *Mol. Phys.* **1983**, *50*, 1055–1076.

[52] Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.* **1997**, *18*, 1463–1472.

[53] Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: An N log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089.

[54] Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A smooth particle mesh Ewald method. *J. Chem. Phys.* **1995**, *103*, 8577–8593.

[55] Hub, J. S.; de Groot, B. L.; van der Spoel, D. g_wham—A Free Weighted Histogram Analysis Implementation Including Robust Error and Autocorrelation Estimates. *J. Chem. Theory Comput.* **2010**, *6*, 3713–3720.

[56] Chen, Y.; Roux, B. Constant-pH Hybrid Nonequilibrium Molecular Dynamics–Monte Carlo Simulation Method. *J. Chem. Theory Comput.* **2015**, *11*, 3919–3931.

[57] Baptista, A. M.; Martel, P. J.; Petersen, S. B. Simulation of protein conformational freedom as a function of pH: constant-pH molecular dynamics using implicit titration. *Proteins: Struct., Funct., Genet.* **1997**, *27*, 523–544.

[58] Baptista, A. M.; Teixeira, V. H.; Soares, C. M. Constant-pH molecular dynamics using stochastic titration. *J. Chem. Phys.* **2002**, *117*, 4184–4200.

[59] Lee, M. S.; Salsbury, F. R.; Brooks, C. L. Constant-pH molecular dynamics using continuous titration coordinates. *Proteins: Struct., Funct., Bioinf.* **2004**, *56*, 738–752.

[60] Mongan, J.; Case, D. A.; McCammon, J. A. Constant pH molecular dynamics in generalized Born implicit solvent. *J. Comput. Chem.* **2004**, *25*, 2038–2048.

[61] Khandogin, J.; Brooks, C. L. Constant pH Molecular Dynamics with Proton Tautomerism. *Biophys. J.* **2005**, *89*, 141–157.

[62] Stern, H. A. Molecular simulation with variable protonation states at constant pH. *J. Chem. Phys.* **2007**, *126*, 164112.

[63] Donnini, S.; Tegeler, F.; Groenhof, G.; Grubmuller, H. Constant pH Molecular Dynamics in Explicit Solvent with λ-Dynamics. *J. Chem. Theory Comput.* **2011**, *7*, 1962–1978.

[64] Lee, J.; Miller, B. T.; Damjanović, A.; Brooks, B. R. Constant pH Molecular Dynamics in Explicit Solvent with Enveloping Distribution Sampling and Hamiltonian Exchange. *J. Chem. Theory Comput.* **2014**, *10*, 2738–2750.

[65] Salomon-Ferrer, R.; Case, D. A.; Walker, R. C. An overview of the Amber biomolecular simulation package. *WIREs Comput. Mol. Sci.* **2013**, *3*, 198–210.

[66] Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kalé, L.; Schulten, K. Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **2005**, *26*, 1781–802.

[67] 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., Bioinf.* **2006**, *65*, 712–725.

[68] Case, D. et al. Amber 14. **2014**, 826.

[69] Onufriev, A.; Bashford, D.; Case, D. A. Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins: Struct., Funct., Bioinf.* **2004**, *55*, 383–394.

[70] Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **1977**, *23*, 327–341.

[71] Miyamoto, S.; Kollman, P. A. Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J. Comput. Chem.* **1992**, *13*, 952–962.

[72] Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual molecular dynamics. *J. Mol. Graphics* **1996**, *14*, 33–38.

[73] Gajiwala, K. S. et al. KIT kinase mutants show unique mechanisms of drug resistance to imatinib and sunitinib in gastrointestinal stromal tumor patients. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 1542–1547.

[74] Liang, L.; Yan, X.-E.; Yin, Y.; Yun, C.-H. Structural and biochemical studies of the PDGFRA kinase domain. *Biochem. Biophys. Res. Commun.* **2016**, *477*, 667–672.

[75] Mol, C. D.; Dougan, D. R.; Schneider, T. R.; Skene, R. J.; Kraus, M. L.; Scheibe, D. N.; Snell, G. P.; Zou, H.; Sang, B. C.; Wilson, K. P. Structural basis for the autoinhibition and STI-571 inhibition of c-Kit tyrosine kinase. *J. Biol. Chem.* **2004**, *279*, 31655–31663.

[76] Meng, Y.; Lin, Y. L.; Roux, B. Computational study of the "DFG-Flip" conformational transition in c-Abl and c-Src tyrosine kinases. *J. Phys. Chem. B* **2015**, *119*, 1443–1456.

[77] Snyder, G. H.; Cennerazzo, M. J.; Karalis, A. J.; Locey, D. Electrostatic influence of local cysteine environments on disulfide exchange kinetics. *Biochemistry* **1981**, *20*, 6509–6519.

[78] Britto, P. J.; Knipling, L.; Wolff, J. The local electrostatic environment determines cysteine reactivity of tubulin. *J. Biol. Chem.* **2002**, *277*, 29018–29027.

[79] Isom, D. G.; Castañeda, C. A.; Cannon, B. R.; García-Moreno, B. Large shifts in pKa values of lysine residues buried inside a protein. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 5260–5.

[80] Pettinger, J.; Jones, K.; Cheeseman, M. D. Lysine-Targeting Covalent Inhibitors. *Angew. Chem. Int. Ed.* **2017**, *56*, 15200–15209.

[81] Paulsen, C. E.; Carroll, K. S. Cysteine-mediated redox signaling: Chemistry, biology, and tools for discovery. *Chem. Rev.* **2013**, *113*, 4633–4679.

[82] Singh, J.; Petter, R. C.; Kluge, A. F. Targeted covalent drugs of the kinase family. *Curr. Opin. Chem. Biol.* **2010**, *14*, 475–480.

[83] Roskoski, R. Ibrutinib inhibition of Bruton protein-tyrosine kinase (BTK) in the treatment of B cell neoplasms. *Pharmacol. Res.* **2016**, *113*, 395–408.

[84] Solca, F.; Dahl, G.; Zoephel, A.; Bader, G.; Sanderson, M.; Klein, C.; Kraemer, O.; Himmelsbach, F.; Haaksma, E.; Adolf, G. R. Target Binding Properties and Cellular Activity of Afatinib (BIBW 2992), an Irreversible ErbB Family Blocker. *J. Pharmacol. Exp. Ther.* **2012**, *343*, 342–350.

[85] Burstein, H. J. et al. Neratinib, an Irreversible ErbB Receptor Tyrosine Kinase Inhibitor, in Patients With Advanced ErbB2-Positive Breast Cancer. *J. Clin. Oncol.* **2010**, *28*, 1301–1307.

[86] Soria, J.-C. et al. Osimertinib in Untreated EGFR -Mutated Advanced Non–Small-Cell Lung Cancer. *N. Engl. J. Med.* **2018**, *378*, 113–125.

[87] Truong, T. H.; Ung, P. M.-U.; Palde, P. B.; Paulsen, C. E.; Schlessinger, A.; Carroll, K. S. Molecular Basis for Redox Activation of Epidermal Growth Factor Receptor Kinase. *Cell Chem. Biol.* **2016**, *23*, 837–848.

[88] Kortemme, T.; Creighton, T. E. Ionisation of Cysteine Residues at the Termini of Model α-Helical Peptides. Relevance to Unusual Thiol pKa Values in Proteins of the Thioredoxin Family. *J. Mol. Biol.* **1995**, *253*, 799–812.

[89] Miranda, J. J. L. Position-dependent interactions between cysteine residues and the helix dipole. *Protein Sci.* **2003**, *12*, 73–81.

[90] Lopes, P. E. M.; Huang, J.; Shim, J.; Luo, Y.; Li, H.; Roux, B.; MacKerell, A. D. Polarizable Force Field for Peptides and Proteins Based on the Classical Drude Oscillator. *J. Chem. Theory Comput.* **2013**, *9*, 5430–5449.

[91] Huang, J.; MacKerell, A. D. Induction of Peptide Bond Dipoles Drives Cooperative Helix Formation in the (AAQAA)3 Peptide. *Biophys. J.* **2014**, *107*, 991–997.

[92] Wood, E. R. et al. 6-Ethynylthieno[3,2-d]- and 6-ethynylthieno[2,3-d]pyrimidin-4-anilines as tunable covalent modifiers of ErbB kinases. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 2773–2778.

[93] Capoferri, L.; Lodola, A.; Rivara, S.; Mor, M. Quantum Mechanics/Molecular Mechanics Modeling of Covalent Addition between EGFR–Cysteine 797 and N -(4-Anilinoquinazolin-6-yl) Acrylamide. *J. Chem. Inf. Model.* **2015**, *55*, 589–599.

[94] Onufriev, A. V.; Alexov, E. Protonation and pK changes in protein–ligand binding. *Q. Rev. Biophys.* **2013**, *46*, 181–209.

[95] Anandakrishnan, R.; Aguilar, B.; Onufriev, A. V. H++ 3.0: Automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. *Nucleic Acids Res.* **2012**, *40*, 537–541.

[96] Olsson, M. H. M.; Søndergaard, C. R.; Rostkowski, M.; Jensen, J. H. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions. *J. Chem. Theory Comput.* **2011**, *7*, 525–537.

" ... A molecular system ... (passes) ... from one state of equilibrium to another ... by means of all possible intermediate paths, but the path most economical of energy will be the more often traveled."

— Henry Eyring, 1945

5

Calculating the Full Free Energy Profile for Covalent Modification of a Kinase Target

Contents

5.1 Abstract

Covalent-binding drugs are experiencing a resurgence in drug discovery and development, owing to their benefits of improved potency and sustained target engagement. These drugs bind to their targets by forming a chemical bond with a nucleophilic residue in the target. The mechanism of binding of a covalent drug consists of both covalent and non-covalent binding steps, with different free energy contributions. Mapping and quantifying these free energy contributions require different computer modelling techniques due to the chemical bond formed. Although, there have been multiple studies reporting on the binding affinity and kinetics of covalent-binding inhibitors, no study to our knowledge has provided a rigorous thermodynamic dissection of the various free energy contributions affecting the covalent and non-covalent binding steps of the chemical process. To address this issue, we employ advanced molecular dynamics and quantum chemical calculations to quantify and describe all the steps involved in the covalent modification of a druggable cysteine in a clinically-validated protein kinase target. The enzyme target investigated is Bruton's tyrosine kinase which is implicated in various forms of leukemia. We model the addition reaction of a potent and selective cyanoacrylamide ligand binding to this target. These calculations provide a rigorous and complete free energy profile of the binding process. Our results indicate that the covalent binding step of the ligand and target is a critical step in the chemical reaction and constitutes a large component of the total binding free energy. Non-covalent interactions between the ligand and individual amino acid residues in the binding pocket of the enzyme are also essential for ligand binding, particularly, van der Waals dispersion forces. These results indicate that the mechanism of covalent modification of a protein occurs through a complex series of steps and that entropy, conformational flexibility, non-covalent interactions, and the formation of covalent linkage are all significant factors in the ultimate binding affinity of a covalent drug for its target.

5.2 Introduction

Drugs that bind to their targets covalently are gaining traction in recent drug development efforts,[1,2] particularly in the burgeoning field of kinase inhibitor drug discovery.[3–5] Despite earlier safety concerns of the pharmaceutical industry about their potential for idiosyncratic adverse events and off-target toxicities,[6] covalent drugs have proven safe and effective in treating diverse clinical indications, including cancer.[7] Also, covalent drugs offer the unique benefits of prolonged duration of therapeutic action, improved efficacy, and high target selectivity.[8] To date, there are over 40 FDA-approved covalent drugs.[9] These include early successful drugs like aspirin and penicillin that covalently-modify active site serine residues in their enzyme targets. More recent examples of successful covalent drugs include cysteine-targeting kinase inhibitors, afatinib and ibrutinib. Afatinib is used in the treatment of metastatic non-small cell lung cancer,[10] and ibrutinib is used for treating B cell cancers like chronic lymphocytic leukemia.[11]

The mechanism of targeted covalent inhibition of a druggable target usually involves the reaction of a nucleophilic moiety of a target protein with an electrophilic functional group of the drug (a.k.a., the "warhead"). This reaction is the central process that governs the rate of inhibition of covalent drugs. Selectivity of a covalent drug for an enzyme target is particularly due to favourable non-covalent interactions plus non-conserved complementarity between the nucleophilic target of the enzyme and warhead of the bound drug. Covalent inhibition is preceded by a non-covalent binding step (Scheme 5.1). This positions the electrophilic warhead of the drug where it can undergo a reaction with the nucleophilic residue of the target enzyme.

Scheme 5.1: Mechanism of covalent inhibition of enzyme (**E**) by inhibitor (**I**). **E**·**I** denotes the non-covalent complex while **E**–**I** signifies the covalent adduct upon chemical reaction.

There are a broad range of chemical motifs in existing drugs[12] that can covalently-modify nucleophilic groups in enzyme targets (e.g., –SH side chain of cysteine). Among these, α,β-unsaturated carbonyl compounds,[13] particularly acrylamides, dominate the current pharmacopeia of covalent inhibitors. These acrylamide-based covalent inhibitors undergo a thiol-Michael addition reaction[14] with a non-catalytic cysteine residue of a desired target (Figure 5.1).[15] Although the mechanism of this

covalent engagement is irreversible, recent work by Taunton and coworkers have shown that these irreversible acrylamide-based inhibitors can be chemically tuned to react with target cysteines in a reversible manner.[16–18] More specifically, they demonstrated that by perturbing the steric and electronic environment around an acrylamide electrophile, reversible covalent inhibitors with remarkably slow off-rates can be synthesized.[18] This is particularly advantageous in that concerns about potential off-target modification of proteins that have been associated with toxicity and immunogenicity can be mitigated due to the reversibility of the covalent binding process. This concept of reversible covalent modification has been explored in designing reversible cysteine-targeting covalent inhibitors of the protein kinase family of enzymes[16,18,19]—one of the most important drug targets of the 21st century.[20]

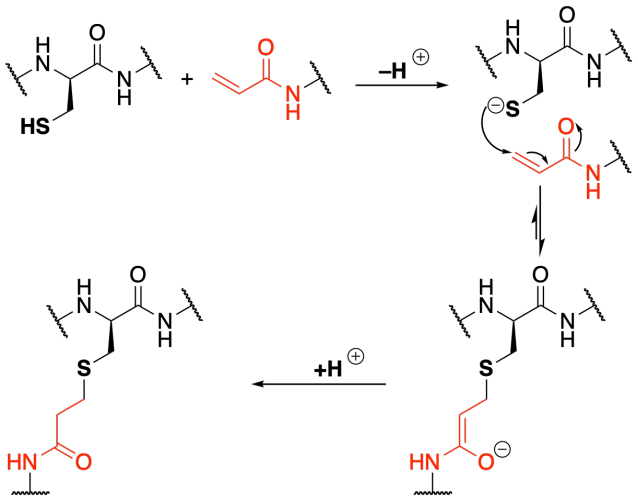

Figure 5.1: Thiol-Michael addition reaction showing the covalent modification of nucleophilic cysteine thiol group of an enzyme target by an acrylamide moiety (in red).

Molecular modeling and quantum chemical methods can provide a wealth of useful insight into the thermodynamics and kinetics of the covalent binding process.[21] Computational methods such as free energy perturbation molecular dynamics (FEP/MD) have been shown to provide accurate estimates of the binding affinity, specificity, and selectivity of ligands for non-covalent enzyme targets.[22–30] Hybrid quantum mechanics/molecular mechanics (QM/MM) methods can be used to describe the potential energy surfaces and reaction thermodynamics of the steps in the covalent reaction mechanism.[31] In particular, the chemical bond formed during the binding step of covalent modification process can only be described using QM methods. Altogether,

these methods when combined can reveal critical molecular information about the relative free energies and barriers associated with the covalent modification mechanism that may be difficult to access experimentally.

Although a number of studies have reported on the kinetics of cysteine-targeted thiol-Michael additions,[32,33] very little attention has been given to understanding in detail the thermodynamic contributions that affect the binding process. This will require calculating the total binding affinity and evaluating the free energy profile of the chemical reaction. To address this issue, we employ alchemical free energy perturbation and hybrid QM/MM molecular dynamics to model all the steps involved in the covalent modification process. Bruton's tyrosine kinase (BTK) is a clinically-validated and attractive target in drug discovery for treating B cell malignancies. Several chemotherapy drugs have been developed that target this kinase enzyme, most notably ibrutinib,[11] which is used for treating various forms of leukemia. BTK contains a druggable cysteine (Cys481) that is accessible for covalent modification by inhibitors. Our model system consists of a protein–ligand complex of BTK with a t-butyl cyanoacrylamide ligand bearing a piperidine linker and pyrazolopyrimidine scaffold, Figure 5.2a. A high-resolution X-ray crystallographic structure has been reported for this model system (PDB ID: 4YHF). The cyanoacrylamide ligand is a highly potent and selective reversible covalent inhibitor of BTK, (Figure 5.2b).[18]

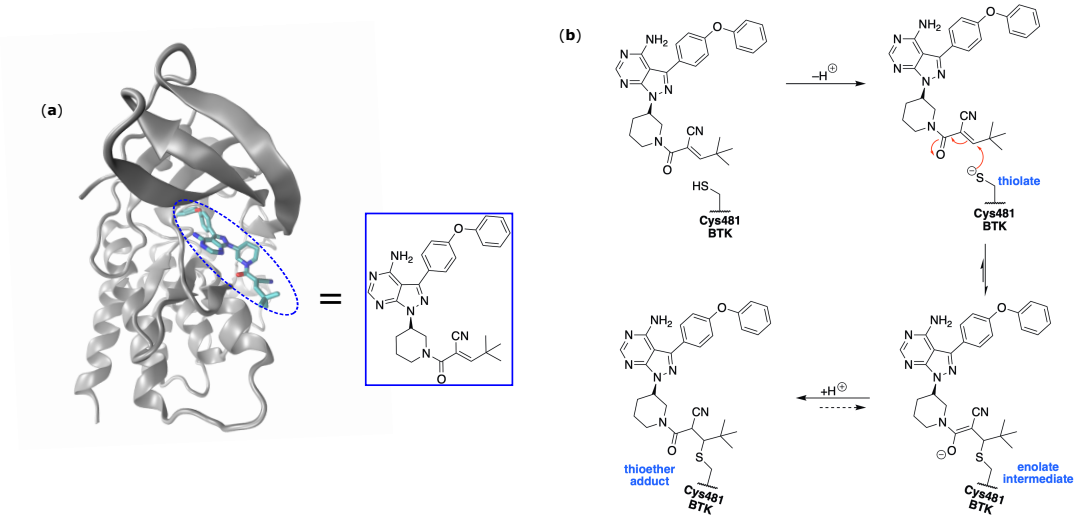

Figure 5.2: (a) Cocrystal structure of BTK complexed with t-butyl cyanoacrylamide inhibitor (PDB ID: 4YHF). The chemical structure of the inhibitor is highlighted. (b) Reaction mechanism showing the steps involved in the reversible covalent modification process for addition of the t-butyl cyanoacrylamide inhibitor to Cys481 of BTK.

To quantify and describe the energetic determinants of all the steps involved in the cysteine-targeting thiol-Michael addition reaction, we performed detailed all-atom FEP/MD simulations and QM/MM MD simulations to calculate a rigorous, complete binding energy profile of the cyanoacrylamide ligand binding to BTK in an explicit aqueous solvent. This combined computational methodology allows for an in-depth thermodynamic dissection of the molecular determinants governing the mechanism of covalent modification.

5.3 Theory & Methods

5.3.1 Ligand–Protein System Setup

The initial structure of BTK complexed with t-butyl cyanoacrylamide ligand was taken from PDB entry 4YHF.[18] MD simulations were performed in explicit solvent and the TIP3P[34] water model was chosen to describe water molecules. The GAAMP[35] method was used to obtain the ligand parameters and the CHARMM36 all-atom protein force field[36] was used to model the protein. Our previous biomolecular simulation studies have shown that the CHARMM36 force field provides the most accurate prediction of the experimental hydration structure of model cysteine thiolates,[37] as well as cysteine pK_a's in proteins[38] and kinase enzymes.[15] All crystallographic resolved water molecules in the structure were retained in the model system. The initial protein–ligand complex was solvated in a simulation cell with dimensions of $72\times72\times80$ Å^3. Sodium ions were added to neutralize the system. All MD simulations of the protein–ligand complex were performed using NAMD 2.13[39] under periodic boundary conditions. A constant temperature of 298.15 K and pressure of 1 atm was applied to the system using the Langevin dynamics and Langevin piston method, respectively. A Langevin damping coefficient of 1 ps^{-1} was used for propagating dynamics and a timestep of 2 fs was used in all calculations. Long range electrostatics interactions were treated using the particle mesh Ewald (PME) method.[40,41] A cutoff distance of 12 Å was applied to Lennard-Jones interactions. A smoothing function was applied from 10 to 12 Å to smoothly truncate van der Waals forces at the cutoff distance. The SHAKE algorithm[42] was applied to constrain covalent bonds involving hydrogen atoms. The model system was initially energy minimized for 1000 steps to eliminate any steric clashes or structural irregularities that may exist within the protein–ligand

molecular assembly. The system was then equilibrated for 20 ns at constant pressure and temperature conditions (NpT ensemble) of 1 atm and 298.15 K, respectively. The equilibration was performed under harmonic restraints using the collective variable (colvars) module[43] in NAMD. This was done in order to ensure that the configuration of the protein–ligand complex remained close to the crystallographic structure. The coordinates of the equilibrated protein–ligand complex were used as a starting structure for the free energy perturbation calculations.

5.3.2 Absolute Binding Free Energy Calculations

To calculate the absolute binding free energy of t-butyl cyanoacrylamide ligand to BTK, we applied alchemical free energy perturbation/lambda-exchange molecular dynamics (FEP/λ-REMD) and umbrella sampling/replica-exchange molecular dynamics (US/REMD) with restraining potentials using the double decoupling protocol developed by Roux and coworkers.[44] This method provides a rigorous step-by-step formulation for computing absolute protein–ligand binding free energies[45–47] and allows for the inclusion of explicit solvent, conformational entropy, and flexibility to yield absolute binding free energies. The alchemical FEP/λ-REMD and US/REMD simulation techniques have achieved good statistical convergence for calculating the absolute binding affinities of small molecule ligands to tyrosine kinases.[22–24] The approach relies on a series of alchemical transformations to compute the absolute binding energy of a ligand to a protein. This follows a step-by-step reversible work staging procedure where the ligand is restrained in its native conformation in the bound state and is then decoupled from its environment. In the decoupling step, the electrostatics and Lennard-Jones interactions between the ligand and its environment (protein and bulk solvent) are gradually switched off. The absolute free energy of ligand binding is simply derived from separate energy contributions from a series of rigorous sequential equilibrium simulations and correspond to the forces and intermolecular interactions of the ligand when in bulk solution and in the protein binding site, Figure 5.3.

The absolute binding free energy ($\Delta G^{o}_{binding}$) of transferring a ligand from bulk solvent (bulk) to a protein binding site (site) can be formally expressed as:[23]

$$\Delta G^{o}_{binding} = \Delta\Delta G^{bulk\rightarrow site}_{conf} + \Delta\Delta G^{bulk\rightarrow site}_{t+r} + \Delta\Delta G^{bulk\rightarrow site}_{int} \tag{5.1}$$

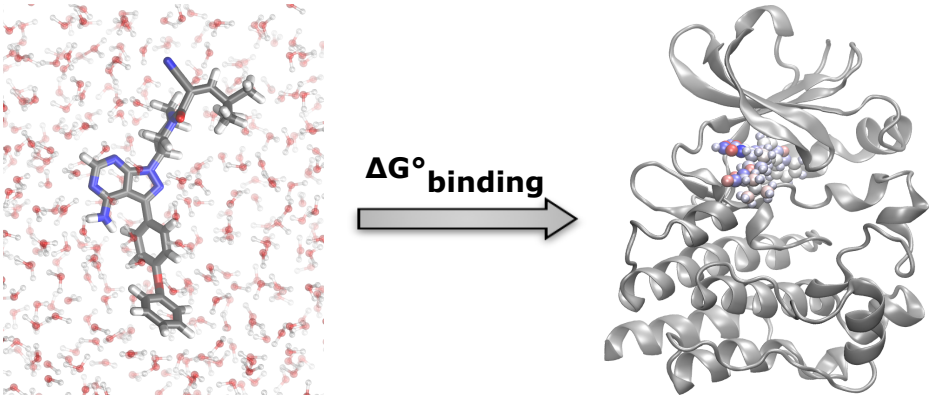

Figure 5.3: Structure of cyanoacrylamide ligand in bulk solution and in BTK binding site. The absolute ligand binding energy is the total free energy of transferring a ligand from aqueous solvent to receptor.

where $\Delta\Delta G_{conf}^{bulk\to site}$ is the conformational free energy cost associated with applying conformational restraints on the ligand in bulk solvent to restrain its conformation near its bound-state and releasing the restraint in the binding site; $\Delta\Delta G_{t+r}^{bulk\to site}$ is the free energy cost of imposing and releasing translational and rotational restraints on the ligand in bulk solvent and in the binding site; $\Delta\Delta G_{int}^{bulk\to site}$ is the difference in free energy of dissociation of the ligand from bulk solvent and its association in the protein binding site that results from intermolecular interactions between the ligand and its environment. All the free energy terms comprising the binding energy in Eqn. (5.1) can further be decomposed into separate individual energy components in bulk solvent and in the protein binding site.

The conformational free energy cost of the ligand upon binding $\Delta\Delta G_{conf}^{bulk\to site}$ is given as:

$$\Delta\Delta G_{conf}^{bulk\to site} = \Delta G_{conf}^{bulk} - \Delta G_{conf}^{site} \tag{5.2}$$

These terms correspond to the free energy contributions associated with applying and releasing conformational restraints of the ligand in bulk solution and in the binding site relative to a reference conformation. US/REMD simulations were used to obtain the conformational free energy of the ligand upon binding. These simulations were carried out by computing the potential of mean force (PMF) of the ligand as a function of the root mean-square deviation (RMSD) relative to its bound-state conformation both in bulk solvent and in the protein binding site. The force constant for all RMSD

restraints used in the simulations was 50 kcal/mol·Å^2. The equilibrated conformation of the bound-state ligand to BTK was used as the reference structure for the RMSD restraint. The US/REMD simulations consisted of eight replicas centered on RMSD offsets increasing from 0.0–3.5 Å in increments of 0.5 Å. Each replica window was sampled for 20 ns, totaling 160 ns of sampling. The conformational free energies of the ligand in bulk solvent (ΔG^{bulk}_{conf}) and in the binding site (ΔG^{site}_{conf}) were calculated from the US/REMD data collected using the weighted histogram analysis method (WHAM),[48] after sorting replica trajectories into their respective files.

The translational and rotational degrees of freedom of the ligand upon binding is given by:

$$\Delta\Delta G^{bulk\rightarrow site}_{t+r} = [-k_B T \ln(F_t C^\circ) - \Delta G^{site}_t] + [-k_B T \ln(F_r) - \Delta G^{site}_r] \tag{5.3}$$

where F_t and F_r denote the translational and rotational factors that are calculated numerically to define the position and orientation of the bound ligand.[44] C° is the standard reference concentration of 1 mol/L or 1/1661 Å^{-3} and T is the absolute temperature in Kelvin. The relative position and orientation of the bound ligand in the complex was described using six internal coordinates; one distance, two angles, and three dihedrals. These values were calculated from the average equilibration trajectory. The force constant for the distance was set to 10 kcal/mol·Å^2, while those of the angles and dihedrals were set to 0.1 kcal/mol·degrees^2. The Gibbs energy associated with imposing each restraint was calculated using thermodynamic integration in the colvars module of NAMD. The free energy contribution due to the translational and rotational restraints on the ligand in the binding site (ΔG^{site}_t and ΔG^{site}_r) were then determined by integrating the gradient profile. The simulation length was 67.2 ns.

The interaction free energy term ($\Delta\Delta G^{bulk\rightarrow site}_{int}$) is decomposed into repulsive, dispersive, and electrostatic contributions, Eqn. (5.4).

$$\Delta\Delta G^{bulk\rightarrow site}_{int} = \Delta\Delta G_{rep} + \Delta\Delta G_{dis} + \Delta\Delta G_{elec} \tag{5.4}$$

where $\Delta\Delta G_{rep} = \Delta\Delta G^{site}_{rep} - \Delta\Delta G^{bulk}_{rep}$, $\Delta\Delta G_{dis} = \Delta\Delta G^{site}_{dis} - \Delta\Delta G^{bulk}_{dis}$, and $\Delta\Delta G_{elec} = \Delta\Delta G^{site}_{elec} - \Delta\Delta G^{bulk}_{elec}$. These terms describe the repulsive, dispersive, and electrostatic

interactions, respectively, of removing the ligand from bulk solvent and inserting it into the binding site. The repulsive and dispersive components of the free energies were obtained from the 6–12 Lennard-Jones potential using the Weeks–Chandler–Anderson[49] decoupling scheme.[50] Alchemical FEP/λ-REMD simulations, staged by three thermodynamic coupling parameters (λ_{rep}, λ_{dis}, λ_{elec}), were applied to compute the ligand interaction free energies in bulk solvent and in the protein binding site. A total number of 36 replicas ($12\lambda_{rep}$, $12\lambda_{dis}$, $12\lambda_{elec}$) were used and the simulation length for each replica window was 20 ns, totaling 720 ns for a complete single run. The simulations were performed in triplicate yielding a total simulation time of 2.16 μs. Interaction free energies separated into their repulsive, dispersive, and electrostatic components for the ligand in bulk solution ($\Delta\Delta G^{bulk}_{rep}$, $\Delta\Delta G^{bulk}_{dis}$, $\Delta\Delta G^{bulk}_{elec}$) and in the binding site ($\Delta\Delta G^{site}_{rep}$, $\Delta\Delta G^{site}_{dis}$, $\Delta\Delta G^{site}_{elec}$) were calculated from the data collected using WHAM. The interaction energies reported are the averages of the three independent replicates.

5.3.3 Potential of Mean Force and Reaction Energies

Hybrid Quantum Mechanics/Molecular Mechanics MD Simulations

We used hybrid QM/MM MD umbrella sampling simulations to calculate the PMF for the addition reaction of the cyanoacrylamide ligand to BTK. The initial structure was taken from the final trajectory of the FEP/λ-REMD simulation. Hybrid QM/MM MD simulations were performed using the comprehensive QM/MM suite implemented in NAMD,[51] with ORCA 4.1.1[52,53] as the QM package. The parameters for the cyanoacrylamide ligand were the same as those used for the MD simulation and were obtained from the GAAMP method. The CHARMM36 all-atom force field was used to describe the protein and TIP3P water model for the explicit solvent water molecules. The QM region defined consisted of the t-butyl cyanoacrylamide warhead with the piperidine linker of the ligand and the thiolate side chain of Cys481 in the protein, Figure 5.4. QM/MM boundary region was treated by using the electrostatic embedding scheme[54] and hydrogen link atoms were used to cap QM regions containing QM–MM bonds.

In order to ensure a stable QM/MM simulation system for the PMF calculations, a series of steps involving QM-based minimization, geometry optimization, and equilibration were performed for the QM and MM regions. The QM region was treated

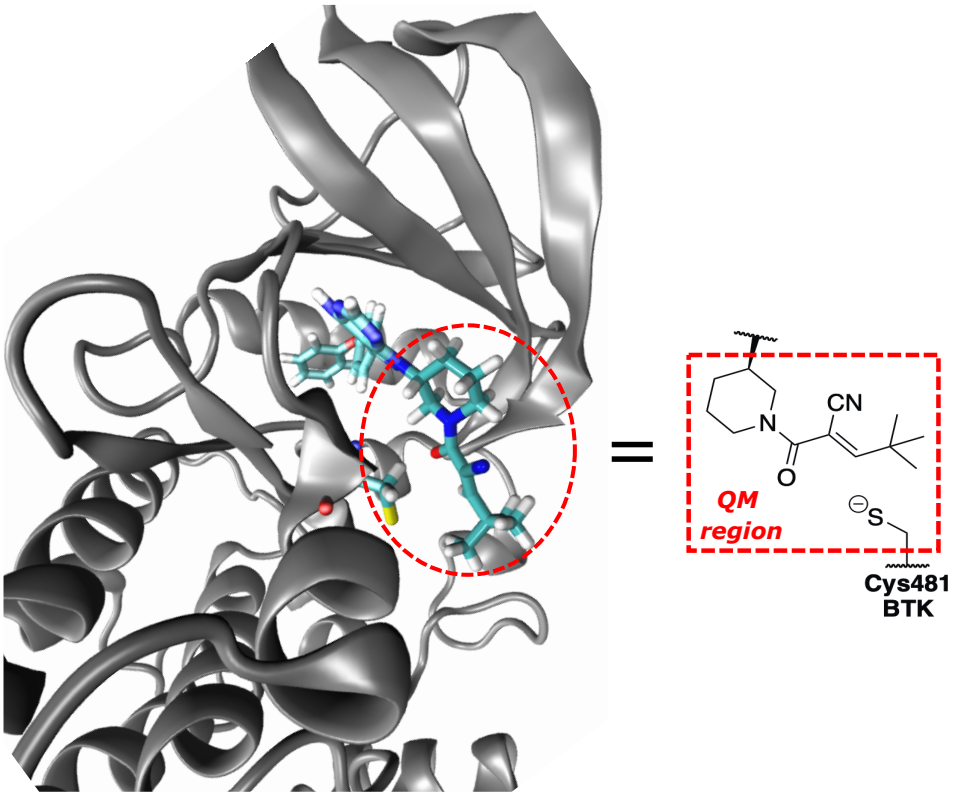

Figure 5.4: The QM region defined in our hybrid QM/MM calculations for the protein–ligand complex.

by the ωB97X-D functional[55] with the def2-TZVP basis set. Grimme's D3 dispersion correction using Becke-Johnson damping function[56] was used for the QM calculations. The ωB97X-D functional performs well for modelling covalent modification of biological thiols,[31] yielding results that are in close agreement with high-level *ab initio* CCSD(T) calculations.[57] The hybrid QM/MM simulation system was subjected to 100 steps of energy minimization, followed by 15 ps equilibration run in the NpT ensemble with temperature maintained at 300 K (using Langevin dynamics) and pressure at 1 atm (using Langevin piston). The PMF was calculated for the interval where the covalent bond is formed between C_β of the t-butyl cyanoacrylamide ligand and the S atom of Cys481 of BTK. This was performed using hybrid QM/MM MD umbrella sampling simulations for the interval r=[1.7, 4.5) Å where the C_β—S bond is formed. The windows for the umbrella sampling simulations were separated by 0.1 Å and a spring constant of 100 kcal/mol·Å^2 was used. For each window, the simulation length was 50 ps, with the first 5 ps discarded as equilibration. The PMF of the C_β—S coordinate was calculated from the umbrella-sampling time series using WHAM.

Quantum Mechanical Calculations with ONIOM

The PMF for the addition reaction of the cyanoacrylamide ligand to BTK receptor yields the enolate intermediate complex as the product of the reaction. In order to attain the final product (i.e., thioether adduct) of the chemical reaction, the α-carbon of the enolate intermediate must be protonated, Figure 5.2. We performed ONIOM[58] QM/MM calculations to compute the relative energy difference between the thioether adduct and enolate intermediate in order to model the final protonation step of the chemical reaction. The ONIOM calculations were performed using Gaussian 16.[59] The final coordinates from the hybrid QM/MM MD equilibration simulation was used as the starting structure for the ONIOM calculations. In order to calculate the Gibbs energy of reaction using this model, it was necessary to develop a simpler model system. A simpler model system of the ligand–protein complex was thus constructed for the ONIOM calculations, Figure 5.5. This truncated model system is based on the ligand interaction diagram from the X-ray crystallographic structure (PDB ID: 4YHF) and considers key interacting amino acid residues within proximal distance from the ligand.

In our ONIOM model, the ligand–protein system is divided into two distinct regions (a.k.a.,"layers"): (1) high-level DFT region, and (2) low-level molecular mechanical region. The high-level region was treated using the ωB97X-D functional with the def2-TZVP basis set, while the low-level layer was treated using the Amber molecular mechanics force field. The high-level layer consisted of the t-butyl cyanoacrylamide warhead with the piperidine linker of the ligand and the Cys481 thiolate side chain of BTK, Figure 5.5. The remainder of the ligand and protein were treated as the low level layer, with the MM charges and parameters the same as those used for the MD simulation. Solvent effects were included in the ONIOM calculations by using the polarized continuum model (PCM).[60,61]

The initial enolate and thioether protein–ligand model complexes were fully optimized in the gas phase using ONIOM (ωB97XD/def2TZVP:AMBER) level of theory. Following this, the model complexes were re-optimized in water to account for solvent effects using PCM. Frequency analyses were then performed on the optimized structures at the same level of theory to verify that there was a true minima, and estimate the thermal correction to the Gibbs energy.

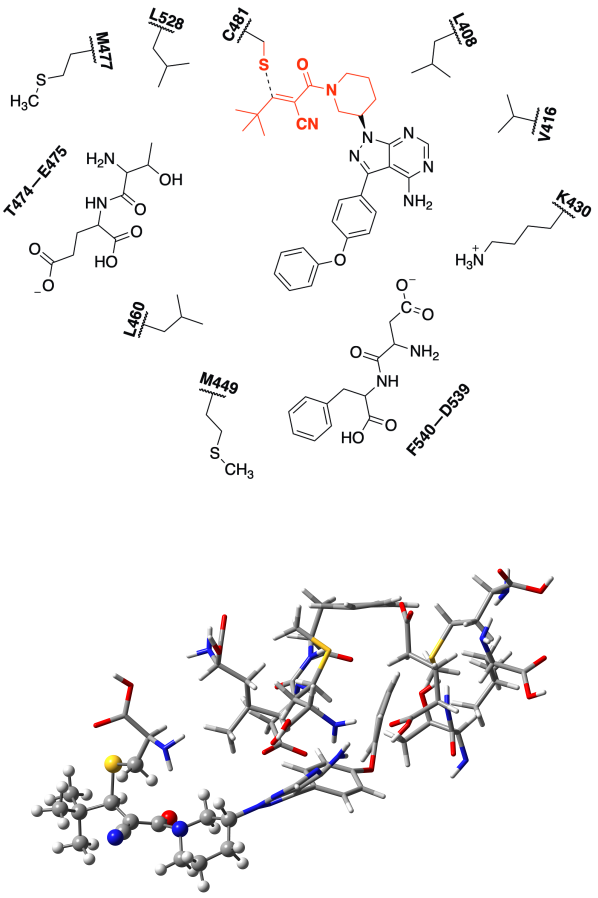

Figure 5.5: 2D (top) and 3D (bottom) representation of ligand–protein model system used for ONIOM calculations. The region constituting high-level layer is indicated in red. For clarity, some residues in the low layer have been omitted in the 3D figure.

5.4 Results and Discussion

Molecular dynamics and quantum chemical methods can be used to calculate the various kinetics and thermodynamic contributions affecting the binding free energy of covalent modification process. In this study, we employ this combined computational approach to evaluate the various thermodynamic contribution terms that affect the reversible covalent modification process of a highly potent and selective inhibitor for BTK. The inhibitor is a t-butyl cyanoacrylamide ligand consisting of a piperidine linker and pyrazolopyrimidine scaffold, Figure 5.2. We computed the absolute free energy of binding for this ligand to BTK and evaluated the free energy profile for the chemical reaction of the covalent modification process.

Alchemical free energy perturbation molecular dynamics and umbrella sampling simulations have been shown to provide accurate estimates of the binding affinity and selectivity of drug-like molecules to biologically relevant enzyme targets,[25,27,28] including protein kinases.[22–24] We calculated the absolute binding affinity of t-butyl cyanoacrylamide ligand to BTK using this method. The computed binding affinity, $\Delta G^{\circ}_{non-covalent}$, represents the non-covalent interactions between the ligand and protein. This takes into account the electrostatic, dispersion, and repulsive free energy contributions driving ligand binding, as well as the translational, rotational, and conformational degrees of freedom underlying the binding process. The covalent binding step of the ligand to the protein kinase enzyme is modelled using hybrid QM/MM molecular dynamics method in aqueous solution. Using this approach, the electrophilic warhead of the cyanoacrylamide ligand and nucleophilic cysteine side chain of the kinase enzyme target are described by quantum mechanics (QM)—which models the chemical reaction step of the covalent modification process. The rest of the ligand and protein are modelled using molecular mechanics (MM) which accounts for the conformational changes, dynamics, and non-covalent binding interactions within the model system. We used this hybrid QM/MM approach to construct a rigorous free energy profile of the ligand binding to BTK.

Overall, these combined computational methods allowed us to model the action of t-butyl cyanoacrylamide ligand binding to BTK in a comprehensive way, through the calculation of both non-covalent and covalent binding free energy terms (i.e., $\Delta G_{non-covalent}$ and $\Delta G_{covalent}$).

5.4.1 Non-covalent Binding Free Energy Contribution

The non-covalent binding free energy of the cyanoacrylamide ligand to BTK was computed thoroughly from a sequence of rigorous equilibrium simulations which characterize the binding/unbinding processes of the ligand in the bound/unbound states. This process yields free energy terms that are separated into various thermodynamic contributions based on the conformation entropies, translational and rotational motions, and the native network of intermolecular interactions underlying the binding process. The absolute non-covalent binding free energy of the ligand to BTK, $\Delta G^{\circ}_{non-covalent}$ which is synonymous to $\Delta G^{\circ}_{binding}$, is determined from the sum of the separate free energy contributions that characterize the reversible association of the ligand to the

protein. These different energy contribution terms correspond to the intermolecular interactions of the ligand following its association and dissociation from the protein kinase receptor. Table 5.1 summarizes the results of the various contributions to the binding free energy of the ligand to BTK.

Table 5.1: Summary of Binding Free Energy Calculations of Ligand to BTK.

$\Delta\Delta G^{bulk \rightarrow site}$ (kcal/mol)	
$\Delta\Delta G_{conf}$	4.4
$\Delta\Delta G_{t+r}$	15.3
$\Delta\Delta G_{elec}$	-5.9
$\Delta\Delta G_{dis}$	-31.9
$\Delta\Delta G_{rep}$	6.7
$\Delta\Delta G_{int}$	-31.1
$\Delta G^{\circ}_{binding}$	-11.4

Each $\Delta\Delta G$ term indicates the free energy difference of the ligand dissociating from bulk solvent and associating in the binding site. The standard deviation of $\Delta G^{\circ}_{binding}$ is 1.3 kcal/mol.

For comparison, the binding free energy of the ligand to BTK was also calculated using the GROMACS molecular dynamics software package[62] with the CHARMM36 and CGenFF[63] force fields, following the protocol reported by Aldeghi *et al.*[27] This approach uses a non-physical thermodynamic cycle to compute absolute binding free energies. The calculated binding free energy of the ligand to BTK using this approach ($\Delta G^{\circ}_{binding} = -11.3 \pm 1.4$ kcal/mol) was found to be in excellent agreement with the binding energy reported in our study ($\Delta G^{\circ}_{binding} = -11.4 \pm 1.3$ kcal/mol; Table 5.1). The agreement of the binding free energy results between these two different approaches makes us confident in our predicted values. Furthermore, the calculated binding free energy result is in good agreement with experimental binding affinity measurements.[18] For example, the experimental inhibitory constant (K_i) value determined for a closely related compound which bears a methylpyrrolidine linker in place of the piperidine linker present in our model cyanoacrylamide ligand is reported as 5.2 nm.[18] This corresponds to a $\Delta G^{\circ}_{exptl.}$ value of -11.30 kcal/mol, which is in good accord with our calculated binding free energy, Table 5.1.

Conformational, Translational and Rotational Free Energies

The free energy contribution arising from the loss of conformational degrees of freedom of the ligand upon binding ($\Delta\Delta G_{conf}^{bulk \rightarrow site}$) is 4.4 kcal/mol. This suggests that the ligand has more conformational freedom in bulk solution than in the protein binding pocket. A plot of the potential of mean force of the ligand in bulk solution and in the protein binding site shows that the ligand adopts a broader range of conformations in solution than when in the protein binding site, Figure 5.6. The potential of mean force is calculated as a function of the root-mean-square deviation relative to a reference bound-state structure. The reference structure was chosen from an equilibrated MD simulation of the ligand–bound X-ray crystallographic structure.

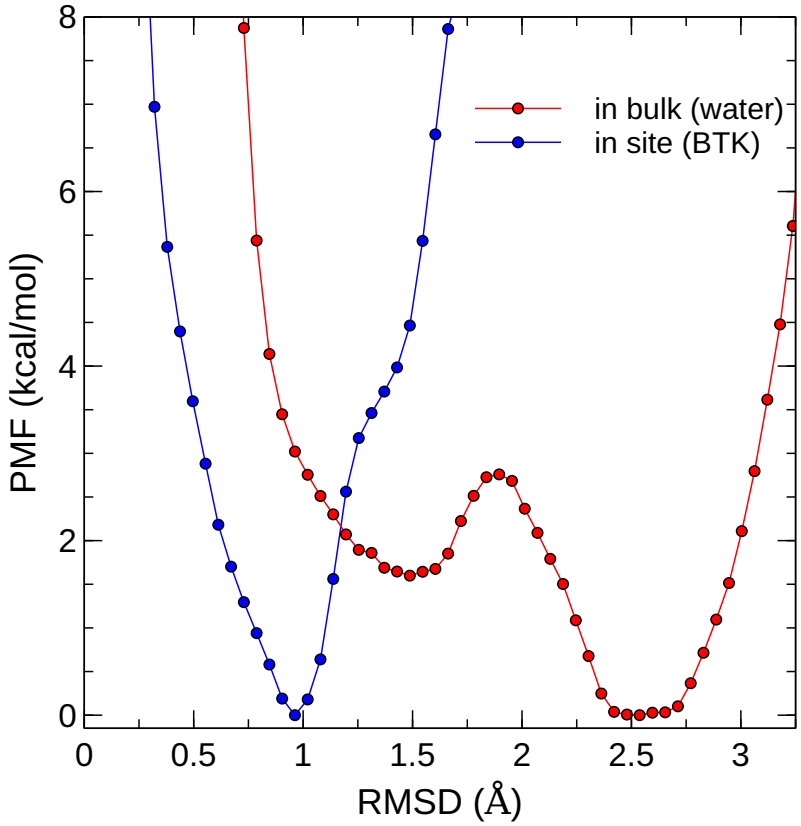

Figure 5.6: Calculated potential of mean force (PMF) of the cyanoacrylamide ligand conformational degrees of freedom as a function of the root-mean-square deviation (RMSD) in the protein binding site (BTK) and in bulk solution (water). The PMF was calculated using US/REMD simulations.

Among the wide range of conformational states that the ligand adopts in bulk solution, the lowest average energy conformations of the ligand is observed at ≈ 2.5 Å, Figure 5.6. On the other hand, a single average conformation is observed for the ligand in the binding pocket of the protein at ≈ 1 Å, Figure 5.6. The PMF of the ligand in bulk solution is also broader than in the protein binding site, a result that confirms the multiple accessible low-energy conformational states that are available to the ligand in bulk solution than in the binding pocket of the protein. Collectively, these PMFs help quantify the loss in conformational entropy of the cyanoacrylamide ligand upon binding to BTK receptor. Figure 5.7 shows the average conformation of the cyanoacrylamide ligand in bulk solution (red) and in the protein binding site (blue) at RMSDs of 2.5 Å and 1.0 Å, respectively

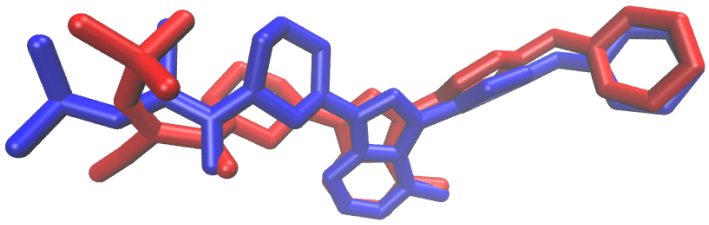

Figure 5.7: Sample conformational states taken up by the cyanoacrylamide ligand in bulk water and in BTK binding site. The red and blue structures represent the average conformational states of ligand in water and binding site at RMSDs of 2.5 Å and 1.0 Å, respectively, relative to the average reference bound-state structure.

The free energy cost arising from restrictions in translational and rotational motions of the ligand upon binding ($\Delta\Delta G_{t+r}^{bulk \rightarrow site}$) is 15.3 kcal/mol. This suggests an even greater entropic penalty for the ligand losing its translational and rotational freedom upon binding to BTK than the penalty arising from the loss of conformational freedom upon binding. On the other hand, the loss of translational and rotational freedom of the ligand upon binding could suggest that this ligand is held more tightly in the binding pocket of the protein by strong non-covalent interactions. Together, the loss of translational, rotational, and conformational degrees of freedom of the ligand accompanying the binding process comes at a considerable free energy cost.

Interaction Free Energies

The interaction free energy ($\Delta\Delta G_{int}^{bulk \rightarrow site}$) associated with the non-covalent dissociation of the ligand from bulk solution and its association in the protein binding site is the largest component of the binding free energy, Table 5.1. Among the three major contributions to the nonbonding interaction energy, the dispersion component is the most dominant ($\Delta\Delta G_{dis}$=−31.9 kcal/mol). This is followed by the electrostatic interactions ($\Delta\Delta G_{elec}$=−5.9 kcal/mol), which is less than one-fifth the magnitude of the dispersion forces. The repulsive contribution opposes the binding of the ligand to the protein kinase ($\Delta\Delta G_{rep}$=6.7 kcal/mol).

These results suggest that the contribution from the dispersion forces play a major role in the non-covalent binding interactions of the cyanoacrylamide ligand to BTK receptor. More specifically, the ligand enjoys more favourable dispersive interactions in the binding pocket of the protein kinase enzyme than in bulk solution. A closer examination of the ligand interaction diagram of the protein–ligand complex shows that the ligand makes strong hydrophobic contacts and van der Waals interactions with amino acid residues in the binding pocket of the protein, Figure 5.8. For example, there exists a favourable $\pi - \pi$ stacking interaction between Phe540 of the protein and the phenyl group of the bound ligand. Additionally, the phenyl group of the ligand makes favourable van der Waals contacts with a host of individual amino acid residues in the binding pocket, including Leu528 and Met449. Hydrogen bonding interactions also contribute to enhancing the non-covalent interactions of the ligand in the binding pocket. One such example is the interaction between the backbone amide hydrogen of Met477 in the protein and a nitrogen in the pyrazolopyrimidine scaffold of the ligand. These non-covalent interactions between the ligand and individual amino acid residues of the protein leads to strong stabilization of the ligand in the binding pocket of BTK receptor, favouring ligand binding.

Although the non-covalent interactions accompanying ligand binding result in unfavourable repulsive interactions, this effect is nullified by the favourable dispersion and electrostatic interactions. The free energy of ligand binding to BTK is dominated by van der Waals dispersion interactions, which reflects a key role of dispersion forces in ligand–protein association. In a similar vein, Roux and coworkers have shown that van der Waals dispersion interactions is primarily responsible for the binding affinity and specificity of the well-known anticancer drug Gleevec to tyrosine kinases.[22–24]

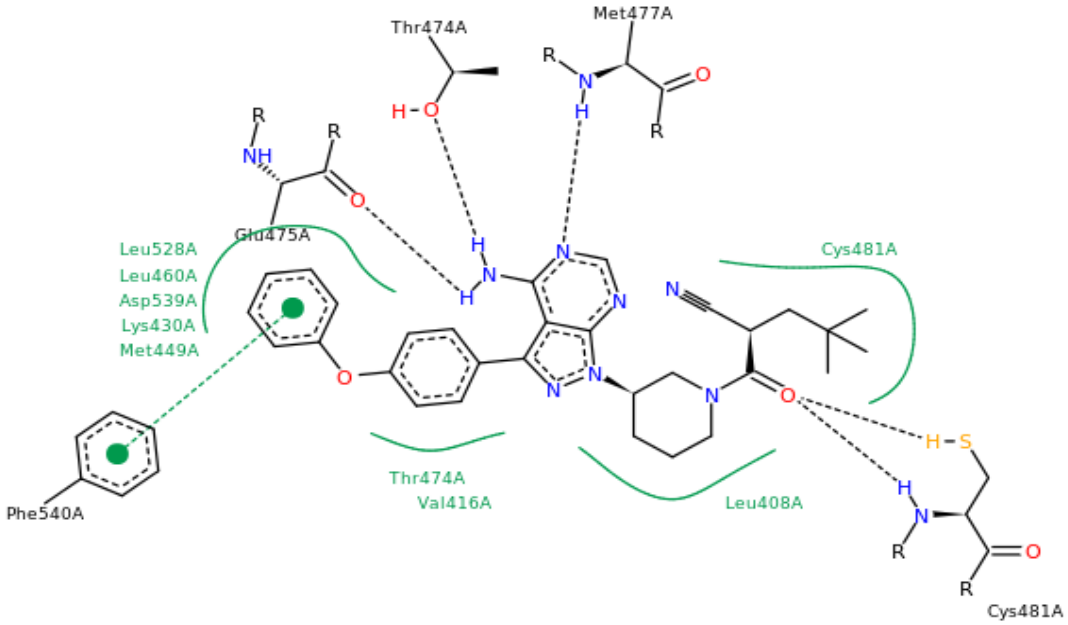

Figure 5.8: Ligand interaction diagram of t-butyl cyanoacrylamide ligand with amino acid residues of BTK (PDB ID: 4YHF). Hydrogen bonding, hydrophobic, and $\pi - \pi$ interactions are indicated by dashed black, solid green, and dashed green lines, respectively.

5.4.2 Covalent Binding Free Energy Contribution

The potential energy surface for the addition reaction of the cyanoacrylamide ligand with BTK in explicit aqueous solution was calculated using QM/MM MD umbrella sampling simulations. ONIOM QM/MM simulations were also performed to compute the relative free energy difference between enolate intermediate and thioether product of the chemical reaction. The ωB97X-D functional which is accurate for modelling covalent modification of biological thiols[31] was used to describe the QM region. The MM region on the other hand, was described using the molecular mechanics force field. Solvent effects were included in the calculations either as individual explicit water molecules or as a dielectric continuum medium. The QM region consisted of the electrophilic cyanoacrylamide warhead of the ligand and Cys481 thiolate side chain of the protein kinase receptor, Figure 5.4. The remaining part of the ligand–protein model system was treated using MM. The PMF was calculated along $r_{C_\beta-S}$ coordinate where the C_β—S bond is formed ($r = 1.7 - 4.5$ Å). Figure 5.9 shows the calculated PMF for the addition reaction of t-butyl cyanoacrylamide ligand with BTK.

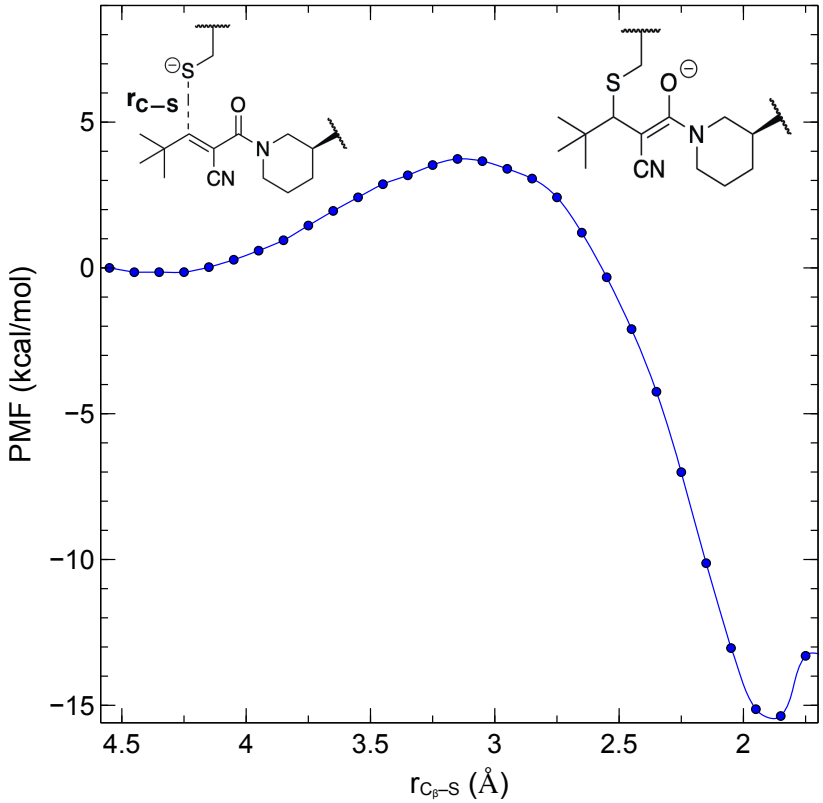

Figure 5.9: PMF for the reaction of t-butyl cyanoacrylamide inhibitor with Cys481 side chain of BTK in aqueous solution. The PMF was calculated using QM/MM MD simulations (QM: ωB97X-D3BJ/def2-TZVP, MM: CHARMM36).

The equilibrium non-covalent ligand:protein complex forms a contact pair at $r_{C_\beta-S}$ distance of approximately 4.45 Å. This complex is used as a reference point for the free energy profile, so its free energy is defined as 0 kcal/mol. Relative to this non-covalent ligand:protein complex, the enolate intermediate is 15.2 kcal/mol more stable, yielding $\Delta G_{enolate} = -15.2$ kcal/mol. There is a modest activation energy barrier of 3.9 kcal/mol (at $r_{C_\beta-S} = 3.15$ Å) for the chemical reaction leading to the formation of the enolate intermediate complex. The free energy minimum for the enolate intermediate is observed at a $C_\beta-S$ distance of 1.85 Å. This distance is slightly more elongated than the typical C−S bond length of approximately 1.8 Å[64] because it is an intermediate state structure.

Using our ONIOM model (Figure 5.5), the computed free energy difference between the enolate intermediate and final thioether product of the chemical reaction was calculated to be -292.9 kcal/mol. This predicted free energy value does

not include the hydration free energy of the proton, which is essential to fully account for the protonation step in the conversion of the enolate intermediate into the thioether product (Figure 5.10). The absolute hydration energy of the proton from high-level, first principles electronic structure calculations is predicted to be -262.4 kcal/mol.[65] Upon taking into account the intrinsic hydration free energy of the proton required for final protonation step of the chemical reaction, the thioether product is found to be 30.5 kcal/mol more stable than the enolate intermediate (i.e., $\Delta G_{thioether} - \Delta G_{enolate} = -30.5$ kcal/mol). This suggests that the thioether product as expected is the thermodynamically favoured product of the chemical reaction.

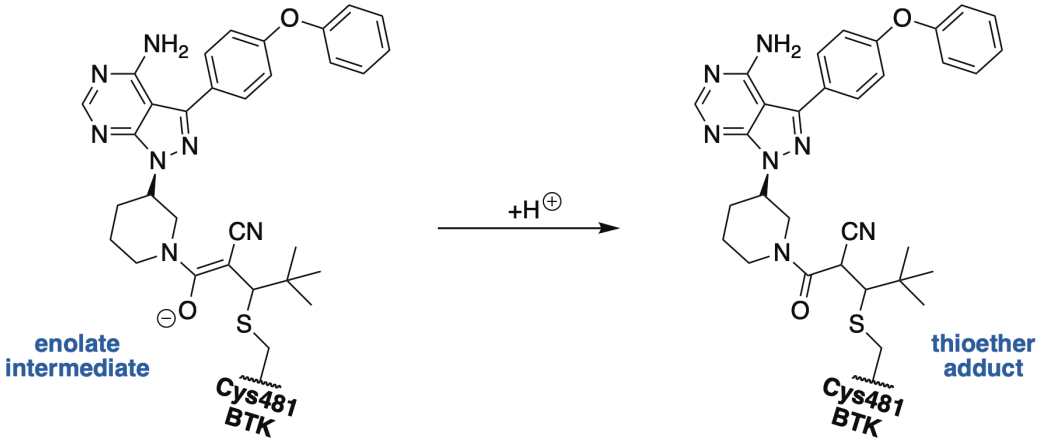

Figure 5.10: Reaction scheme showing the protonation step required for the conversion of ligand–protein enolate intermediate to thioether adduct.

Combining these results, the free energy contribution due to the covalent binding chemical process of the cyanoacrylamide ligand to BTK is -45.7 kcal/mol (i.e., $\Delta G_{covalent} = -45.7$ kcal/mol). This represents a strongly exergonic chemical process. The results suggest that the interaction energy that results from the covalent bond formation contributes the largest share to the total binding affinity.

5.4.3 Free Energy Profile of Covalent Modification

Figure 5.11 illustrates the free energy profile for the addition reaction of BTK with the cyanoacrylamide ligand, which represents a highly exergonic chemical process. This reaction consists of multiple steps involving the target cysteine residue and the ligand. The first step in the reaction involves the deprotonation of the thiol side

chain of Cys481 in BTK. The deprotonation Gibbs energy of Cys481 in BTK is estimated from its calculated pK_a value of 9.47, computed using advanced constant-pH methods following our published protocol.[15] This pK_a value corresponds to a Gibbs energy of 1.2 kcal/mol at 298.15 K. Following this, the next step is the formation of non-covalent protein:ligand complex ($\Delta G_{non-covalent} = -11.4$ kcal/mol). The final step involves a chemical reaction between the thiolate group of the cysteine and electrophilic cyanoacrylamide warhead of the ligand. This reaction step leads to the formation of an enolate intermediate, before leading to the formation of the final thioether covalent complex product. The activation energy required for this chemical process is modest ($\Delta G^{TS}_{binding} = 3.9$ kcal/mol). The enolate intermediate formed from this chemical reaction is ≈15 kcal/mol more stable than the non-covalent state and has an absolute Gibbs energy (i.e., $\Delta G_{enolate}$) of -26.6 kcal/mol. The final covalent complex is approximately 30 kcal/mol more stable than the enolate intermediate.

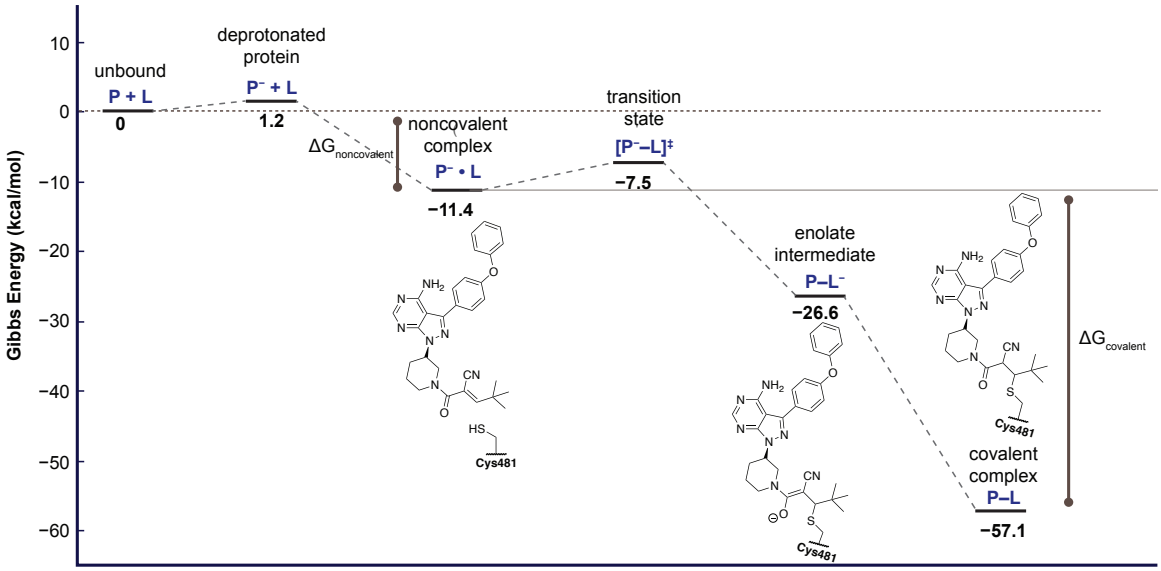

Figure 5.11: Free energy profile of covalent modification of Cys481 in BTK by cyanoacrylamide inhibitor. Cys481 thiol side chain in BTK protein is indicated as P and the cyanoacrylamide ligand is indicated as L.

5.5 Conclusion

In summary, we have employed advanced multiscale simulation methods to model all the steps involved in the covalent modification process of a druggable cysteine in Bruton's tyrosine kinase (BTK) enzyme. BTK is a clinically-validated enzyme target that is of interest in drug discovery, particularly for treating B cell cancers. We explore the covalent modification of Cys481 in this enzyme target by modelling its addition reaction to a t-butyl cyanoacrylamide inhibitor. We quantify the various energetic determinants that contribute to the covalent and non-covalent free energy binding components of the chemical reaction. In addition, a rigorous, complete free energy profile of the inhibitor binding to the enzyme target in aqueous solution is calculated. The results indicate that the chemical reaction leading to the formation of the covalent adduct represents a highly exergonic process on the free energy profile. The covalent binding between C_β of the cyanoacrylamide ligand and Cys481 thiolate side chain of the target represents a critical step in the covalent modification process. This chemical step yields a free energy of -45.7 kcal/mol relative to the non-covalent interaction energy of -31.1 kcal/mol. Van der Waals dispersion forces between the ligand and individual amino acid residues in the protein binding pocket are the largest component of the non-covalent binding energy, and greatly favour ligand binding. The results highlight the importance of both covalent and non-covalent free energy contributions to the thermodynamics of ligand binding processes. It also demonstrates the potential of computer modelling to provide detailed information regarding ligand–protein interactions for drug design and discovery.

Bibliography

[1] Singh, J.; Petter, R. C.; Baillie, T. A.; Whitty, A. The resurgence of covalent drugs. *Nat. Rev. Drug Discov.* **2011**, *10*, 307–317.

[2] Baillie, T. A. Targeted Covalent Inhibitors for Drug Design. *Angew. Chem. Int. Ed.* **2016**, *55*, 13408–13421.

[3] Singh, J.; Petter, R. C.; Kluge, A. F. Targeted covalent drugs of the kinase family. *Curr. Opin. Chem. Biol.* **2010**, *14*, 475–480.

[4] Berndt, N.; Karim, R. M.; Schönbrunn, E. Advances of small molecule targeting of kinases. *Curr. Opin. Chem. Biol.* **2017**, *39*, 126–132.

[5] Ferguson, F. M.; Gray, N. S. Kinase inhibitors: the road ahead. *Nat. Rev. Drug Discovery* **2018**, *17*, 353–377.

[6] Potashman, M. H.; Duggan, M. E. Covalent Modifiers: An Orthogonal Approach to Drug Design. *J. Med. Chem.* **2009**, *52*, 1231–1246.

[7] Zhang, J.; Yang, P. L.; Gray, N. S. Targeting cancer with small molecule kinase inhibitors. *Nat. Rev. Cancer* **2009**, *9*, 28–39.

[8] Bauer, R. A. Covalent inhibitors in drug discovery: from accidental discoveries to avoided liabilities and designed therapies. *Drug Discov. Today* **2015**, *20*, 1061–1073.

[9] Vasudevan, A.; Argiriadi, M. A.; Baranczak, A.; Friedman, M. M.; Gavrilyuk, J.; Hobson, A. D.; Hulce, J. J.; Osman, S.; Wilson, N. S. Covalent binders in drug discovery. *Prog. Med. Chem.* **2019**, *58*, 1.

[10] Joshi, M.; Rizvi, S. M.; Belani, C. P. Afatinib for the treatment of metastatic non-small cell lung cancer. *Cancer Manag. Res.* **2015**, *7*, 75.

[11] Byrd, J. C. et al. Targeting BTK with ibrutinib in relapsed chronic lymphocytic leukemia. *N. Engl. J. Med.* **2013**, *369*, 32–42.

[12] Gehringer, M.; Laufer, S. A. Emerging and re-emerging warheads for targeted covalent inhibitors: applications in medicinal chemistry and chemical biology. *J. Med. Chem.* **2018**, *62*, 5673–5724.

[13] Jackson, P. A.; Widen, J. C.; Harki, D. A.; Brummond, K. M. Covalent Modifiers: A Chemical Perspective on the Reactivity of α,β-Unsaturated Carbonyls with Thiols via Hetero-Michael Addition Reactions. *J. Med. Chem.* **2017**, *60*, 839–885.

[14] Nair, D. P.; Podgorski, M.; Chatani, S.; Gong, T.; Xi, W.; Fenoli, C. R.; Bowman, C. N. The thiol-Michael addition click reaction: a powerful and widely used tool in materials chemistry. *Chem. Mater.* **2013**, *26*, 724–744.

[15] Awoonor-Williams, E.; Rowley, C. N. How Reactive are Druggable Cysteines in Protein Kinases? *J. Chem. Inf. Model.* **2018**, *58*, 1935–1946.

[16] Serafimova, I. M.; Pufall, M. A.; Krishnan, S.; Duda, K.; Cohen, M. S.; Maglathlin, R. L.; McFarland, J. M.; Miller, R. M.; Frödin, M.; Taunton, J. Reversible targeting of noncatalytic cysteines with chemically tuned electrophiles. *Nat. Chem. Biol.* **2012**, *8*, 471–476.

[17] Miller, R. M.; Paavilainen, V. O.; Krishnan, S.; Serafimova, I. M.; Taunton, J. Electrophilic fragment-based design of reversible covalent kinase inhibitors. *J. Am. Chem. Soc.* **2013**, *135*, 5298–5301.

[18] Bradshaw, J. M. et al. Prolonged and tunable residence time using reversible covalent kinase inhibitors. *Nat. Chem. Biol.* **2015**, *11*, 525–531.

[19] London, N.; Miller, R. M.; Krishnan, S.; Uchida, K.; Irwin, J. J.; Eidam, O.; Gibold, L.; Bonnet, R.; Cimermančič, P.; Shoichet, B. K.; Taunton, J. Covalent Docking of Large Libraries for the Discovery of Chemical Probes. *Nat. Chem. Biol.* **2014**, *10*, 1066–1072.

[20] Cohen, P. Protein kinases—the major drug targets of the twenty-first century? *Nat. Rev. Drug Discovery* **2002**, *1*, 309.

[21] Awoonor-Williams, E.; Walsh, A. G.; Rowley, C. N. Modeling covalent-modifier drugs. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* **2017**, *1865*, 1664–1675.

[22] Lin, Y.-L.; Meng, Y.; Jiang, W.; Roux, B. Explaining why Gleevec is a specific and potent inhibitor of Abl kinase. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110*, 1664–1669.

[23] Lin, Y.-L.; Roux, B. Computational analysis of the binding specificity of Gleevec to Abl, c-Kit, Lck, and c-Src tyrosine kinases. *J. Am. Chem. Soc.* **2013**, *135*, 14741–14753.

[24] Lin, Y.-L.; Meng, Y.; Huang, L.; Roux, B. Computational study of Gleevec and G6G reveals molecular determinants of kinase inhibitor selectivity. *J. Am. Chem. Soc.* **2014**, *136*, 14753–14762.

[25] Alsamarah, A.; LaCuran, A. E.; Oelschlaeger, P.; Hao, J.; Luo, Y. Uncovering molecular bases underlying bone morphogenetic protein receptor inhibitor selectivity. *PloS One* **2015**, *10*.

[26] Wang, L. et al. 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*, 2695–2703.

[27] Aldeghi, M.; Heifetz, A.; Bodkin, M. J.; Knapp, S.; Biggin, P. C. Accurate calculation of the absolute free energy of binding for drug molecules. *Chem. Sci.* **2016**, *7*, 207–218.

[28] Aldeghi, M.; Heifetz, A.; Bodkin, M. J.; Knapp, S.; Biggin, P. C. Predictions of ligand selectivity from absolute binding free energy calculations. *J. Am. Chem. Soc.* **2017**, *139*, 946–957.

[29] Chatterjee, P.; Botello-Smith, W. M.; Zhang, H.; Qian, L.; Alsamarah, A.; Kent, D.; Lacroix, J. J.; Baudry, M.; Luo, Y. Can Relative Binding Free Energy Predict Selectivity of Reversible Covalent Inhibitors? *J. Am. Chem. Soc.* **2017**, *139*, 17945–17952.

[30] Moraca, F.; Negri, A.; de Oliveira, C.; Abel, R. Application of Free Energy Perturbation (FEP+) to Understanding Ligand Selectivity: A Case Study to Assess Selectivity Between Pairs of Phosphodiesterases (PDE's). *J. Chem. Inf. Model.* **2019**, *59*, 2729–2740.

[31] Awoonor-Williams, E.; Isley III, W. C.; Dale, S. G.; Johnson, E. R.; Yu, H.; Becke, A. D.; Roux, B.; Rowley, C. N. Quantum Chemical Methods for Modeling Covalent Modification of Biological Thiols. *J. Comput. Chem.* **2020**, *41*, 427–438.

[32] Krenske, E. H.; Petter, R. C.; Houk, K. N. Kinetics and Thermodynamics of Reversible Thiol Additions to Mono- and Diactivated Michael Acceptors: Implications for the Design of Drugs That Bind Covalently to Cysteines. *J. Org. Chem.* **2016**, *81*, 11726–11733.

[33] Krishnan, S.; Miller, R. M.; Tian, B.; Mullins, R. D.; Jacobson, M. P.; Taunton, J. Design of Reversible, Cysteine-Targeted Michael Acceptors Guided by Kinetic and Computational Analysis. *J. Am. Chem. Soc.* **2014**, *136*, 12624–12630.

[34] Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79*, 926–935.

[35] Huang, L.; Roux, B. Automated force field parameterization for nonpolarizable and polarizable atomic models based on ab initio target data. *J. Chem. Theory Comput.* **2013**, *9*, 3543–3556.

[36] Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E. M.; Mittal, J.; Feig, M.; Alexander D. MacKerell, J. Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone ϕ, ψ and Side-Chain $\chi 1$ and $\chi 2$ Dihedral Angles. *J. Chem. Theory Comput.* **2012**, *8*, 3257–3273.

[37] Awoonor-Williams, E.; Rowley, C. N. The hydration structure of methylthiolate from QM/MM molecular dynamics. *J. Chem. Phys.* **2018**, *149*, 045103.

[38] Awoonor-Williams, E.; Rowley, C. N. Evaluation of Methods for the Calculation of the pKa of Cysteine Residues in Proteins. *J. Chem. Theory Comput.* **2016**, *12*, 4662–4673.

[39] Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K. Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **2005**, *26*, 1781–1802.

[40] Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald: An N $\cdot$ log(N) method for Ewald sums in Large Systems. *J. Chem. Phys.* **1993**, *98*, 10089–10092.

[41] Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A smooth particle mesh Ewald method. *J. Chem. Phys.* **1995**, *103*, 8577–8593.

[42] Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **1977**, *23*, 327–341.

[43] Fiorin, G.; Klein, M. L.; Hénin, J. Using collective variables to drive molecular dynamics simulations. *Mol. Phys.* **2013**, *111*, 3345–3362.

[44] Wang, J.; Deng, Y.; Roux, B. Absolute Binding Free Energy Calculations Using Molecular Dynamics Simulations with Restraining Potentials. *Biophys J.* **2006**, *91*, 2798–2814.

[45] Woo, H.-J.; Roux, B. Calculation of absolute protein–ligand binding free energy from computer simulations. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6825–6830.

[46] Deng, Y.; Roux, B. Calculation of standard binding free energies: Aromatic molecules in the T4 lysozyme L99A mutant. *J. Chem. Theory Comput.* **2006**, *2*, 1255–1273.

[47] Deng, Y.; Roux, B. Computations of standard binding free energies with molecular dynamics simulations. *J. Phys. Chem. B* **2009**, *113*, 2234–2246.

[48] Grossfield, A. WHAM: the weighted histogram analysis method, version 2.0. 9. *Available at membrane. urmc. rochester. edu/content/wham. Accessed November* **2013**, *15*, 2013.

[49] Weeks, J. D.; Chandler, D.; Andersen, H. C. Role of repulsive forces in determining the equilibrium structure of simple liquids. *J. Chem. Phys.* **1971**, *54*, 5237–5247.

[50] Deng, Y.; Roux, B. Hydration of amino acid side chains: Nonpolar and electrostatic contributions calculated from staged molecular dynamics free energy simulations with explicit water molecules. *J. Phys. Chem. B* **2004**, *108*, 16567–16576.

[51] Melo, M. C. R.; Bernardi, R. C.; Rudack, T.; Scheurer, M.; Riplinger, C.; Phillips, J. C.; Maia, J. D. C.; Rocha, G. B.; Ribeiro, J. V.; Stone, J. E.; Neese, F.; Schulten, K.; Luthey-Schulten, Z. NAMD goes quantum: an integrative suite for hybrid simulations. *Nat. Methods* **2018**, *15*, 351–354.

[52] Neese, F. The ORCA program system. *WIREs Comput. Mol. Sci.* **2012**, *2*, 73–78.

[53] Neese, F. Software update: the ORCA program system, version 4.0. *WIREs Comput. Mol. Sci.* **2018**, *8*, e1327.

[54] Bakowies, D.; Thiel, W. Hybrid models for combined quantum mechanical and molecular mechanical approaches. *J. Phys. Chem.* **1996**, *100*, 10580–10594.

[55] Chai, J.-D.; Head-Gordon, M. Long-range Corrected Hybrid Density Functionals with Damped Atom-Atom Dispersion Corrections. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6615–6620.

[56] Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the damping function in dispersion corrected density functional theory. *J. Comput. Chem.* **2011**, *32*, 1456–1465.

[57] Smith, J. M.; Jami Alahmadi, Y.; Rowley, C. N. Range-separated DFT functionals are necessary to model Thio-Michael additions. *J. Chem. Theory Comput.* **2013**, *9*, 4860–4865.

[58] Chung, L. W.; Sameera, W.; Ramozzi, R.; Page, A. J.; Hatanaka, M.; Petrova, G. P.; Harris, T. V.; Li, X.; Ke, Z.; Liu, F.; Hai-Bei, L.; Ding, L.; Marokuma, K. The ONIOM method and its applications. *Chem. Rev.* **2015**, *115*, 5678–5796.

[59] Frisch, M. J. et al. Gaussian 16 Revision C.01. 2016; Gaussian Inc. Wallingford CT.

[60] Tomasi, J.; Persico, M. Molecular interactions in solution: an overview of methods based on continuous distributions of the solvent. *Chem. Rev.* **1994**, *94*, 2027–2094.

[61] Cances, E.; Mennucci, B.; Tomasi, J. A new integral equation formalism for the polarizable continuum model: Theoretical background and applications to isotropic and anisotropic dielectrics. *J. Chem. Phys.* **1997**, *107*, 3032–3041.

[62] Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2015**, *1-2*, 19–25.

[63] Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; Mackerell, A. D. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comput. Chem.* **2010**, *31*, 671–690.

[64] Trinajstić, N. Calculation of carbon-sulphur bond lengths. *Tetrahedron Lett.* **1967**, 1529–1532.

[65] Zhan, C.-G.; Dixon, D. A. Absolute hydration free energy of the proton from first-principles electronic structure calculations. *J. Phys. Chem. A* **2001**, *105*, 11534–11540.

"We cannot solve our problems with the same thinking we used when we created them."

— Albert Einstein

6

Summary and Outlook

Contents

6.1 Summary

Computer modelling plays an integral role in the pharmaceutical industry by aiding in drug discovery and development. Covalent drugs, particularly those targeting cysteine residues in enzyme targets, have garnered significant renewed interest in drug discovery owing to their unique benefits of prolonged therapeutic action, improved efficacy, and high target selectivity. The mechanism of inhibition of covalent drugs consists of both covalent and non-covalent binding energy contributions, which require different computational methods to compute. Computer modelling has not been used for modelling covalent-modifier drugs. Furthermore, methods to predict the reactivity of druggable targets in enzymes for covalent modification have received little attention. The fundamental studies presented within this thesis seeks to address some of these research questions. The primary goal of this thesis has been to explore computational methods for modelling all the steps in the covalent modification of targetable cysteine residues in enzyme targets, so as to inform drug design and discovery efforts.

The first step in the covalent modification of a druggable cysteine is the deprotonation of thiol side chain (–SH) to form the more reactive thiolate ($–S^-$). The rate of this reaction is dependent on the acidity or pK_a of the cysteine thiol. Methods for the prediction of cysteine residues in proteins are less established and have received little attention. In Chapter 2, a benchmark assessment of different computational methods was performed in an effort to evaluate their predictive accuracy in calculating experimental cysteine pK_a's. Computational methods that employ both explicit and implicit solvent models for computing cysteine pK_a's were evaluated for a test set of proteins. Results indicated that explicit solvent models are systematically more accurate than implicit solvent models. Among the explicit solvent models, the accuracy of the computed cysteine pK_a's tended to be sensitive to the force field parameters used. In particular, the results from the CHARMM36 force field was more accurate than the Amber force field; RMSD of 2.4 for CHARMM36 force field versus 3.2 for Amber force field. This highlights a limitation of current molecular mechanical force fields in cysteine pK_a calculations. It is possible that reparameterization of existing force fields or the development of new, more sophisticated force fields could yield more accurate results for cysteine pK_a calculations.

In an effort to understand why the CHARMM36 and Amber protein force fields gave different pK_a predictions and to ascertain which is a more realistic model, advanced multiscale computational methods were performed to evaluate the hydration structure of a model cysteine thiolate in aqueous solution (Chapter 3). Hybrid quantum mechanical/molecular mechanical (QM/MM) molecular dynamics simulations and free energy perturbation calculations were used to characterize the solution structure for models of methylthiolate. The results suggested that the CHARMM36 model for thiolate generally provides a better description of the solvation structure and hydration energies of methylthiolate than the Amber force field. This is attributed to the non-bonded parameters present within the different force fields. More specifically, the CHARMM36 force field uses different Lennard-Jones parameters to describe the thiol and thiolate states, while the Amber force field uses the same parameters for the thiol and thiolate states. This study showed that distinct non-bonded parameters are essential in describing the protonated/deprotonated states of model cysteine side chains in biomolecular simulations.

Protein kinases have proven to be major drug targets for treating diverse clinical indications. In fact, they are considered to be one of the most important drug targets of the 21^{st} century.[1] One strategy that is used to target a kinase enzyme implicated in human disease is to covalently-modify a nucleophilic amino acid side chain group in the enzyme by an electrophilic inhibitor. Non-catalytic cysteine residues within the active site region of kinases have been the primary target of this approach. Although a number of studies have used structural analysis to identify cysteine residues in kinases that undergo covalent modification readily, research on the intrinsic reactivity of druggable kinase cysteines have not been reported. Furthermore, few experimental kinase cysteine pK_a's have been determined. In Chapter 4, the reactivity of druggable cysteines in protein kinases was predicted based on their computed pK_a's. Important oncogenic mutants of these kinases were also included in the study. The CHARMM36 explicit solvent method together with other rigorous pK_a calculation methods that are capable of describing variable protonation and pH state effects of ionizable residues in proteins were employed. Results suggested that there is a broad range in the acidity of druggable cysteines in protein kinases, indicating enormous differences in their reactivity towards drug molecules. The general trend observed was that druggable cysteines in kinases have elevated pK_a's and tend to be buried away from contact with solvent water molecules. Inter-residue electrostatic interactions and the degree

of solvation of the active site cysteine thiolate were found to be prime determining factors for the perturbation of kinase cysteine pK_a's.

In Chapter 5, alchemical free energy perturbation and QM/MM molecular dynamics were used to model all the steps in the covalent modification process of Cys481 in Bruton's tyrosine kinase (BTK) by a cyanoacrylamide inhibitor. BTK is a clinically-validated enzyme target for treating B cell cancers. In an effort to quantify and describe the steps in the covalent modification chemical process, the energetic determinants of all the terms contributing to the covalent binding of the ligand to BTK were computed. In addition, a rigorous, complete binding energy profile of the binding process in explicit solvent was also calculated. The results suggest that both non-covalent and covalent binding free energy contributions are important in the covalent modification process, with the latter being the most significant contributor to the total free energy of ligand binding. For the non-covalent binding free energy contribution, van der Waals dispersion forces between the ligand and binding pocket is the most dominant and largest component of the ligand protein interaction energy. The covalent binding step of the addition reaction between BTK and the cyanoacrylamide ligand represents a highly exergonic chemical process.

6.2 Future Directions

The work presented in Chapter 2 highlights the existing barriers and limitations of current computational methods in accurately predicting experimental cysteine pK_a's. Although explicit solvent models have proved to be significantly more accurate than implicit solvent models for cysteine pK_a calculation, there is further room for improvement.

One way to improve the accuracy of these models is by further validation, optimization, and development of existing molecular mechanical force field parameters. This issue is brought to light in Chapter 3 which demonstrated that cysteine pK_a calculations could greatly benefit from improvements in the Lennard-Jones parameters that account for the difference in non-bonded interactions of the thiol/thiolate states of cysteine side chain. Another area of improvement in the accuracy of existing models for cysteine pK_a calculations is the inclusion of polarizable force field. Electrostatic interactions are significant in the perturbation of pK_a's of titratable residues from

their intrinsic solution pK_a's. More significantly, the thiolate ($–S^-$) is a diffuse, polarizable anion and for that reason induced polarization could have a sizable effect on its stability in a protein—influencing the acidity of a target cysteine residue. Further work could include incorporating induced polarization effects in cysteine pK_a calculations by using, for example, the Drude polarizable force field.[2,3] This could improve the accuracy of pK_a calculations over the methods that employ non-polarizable force fields, especially in cases where the charged state of the amino acid strongly polarizes its environment. Additionally, conformational sampling and pH-coupled behavior of amino acid residues is another important area of consideration for improvement in pK_a models. This will allow pK_a's to be predicted more rigorously as shown in Chapter 4, and is particularly important for the quantitative calculation of catalytic cysteine pK_a's whose pK_a's are often coupled to other ionizable residues.

There are relatively few experimentally determined pK_a's for cysteine residues in proteins that have been reported in the literature. This has sparked efforts in using computation to address this issue, which is a fundamental theme of this thesis. Additionally, experimental measurement of more cysteine pK_a's in proteins will allow for a more thorough and comprehensive evaluation of existing pK_a methods. Of particular importance are the pK_a's of noncatalytic residues in protein active sites (Chapter 4), which are typically the target of covalent modifier drugs. Overall, progress in the calculation and prediction of titratable residue pK_a's in proteins will require collaborative efforts from experimentalists and theoreticians alike;[4] redefining the conceptual framework behind the underpinnings of acid-base equilibria in proteins and biomacromolecules.

Lastly, methods for modelling the covalent modification of druggable targets in biologically relevant enzymes are computationally expensive, and require some level of technical expertise (e.g., coding and scripting) to successfully perform such calculations in an accurate way. Although, Chapter 5 presents a successful multiscale approach to quantify the energetics of both the covalent and non-covalent aspects of ligand binding to BTK receptor, this approach requires access to large computational resources which may not be readily available to a user. Additionally, the high cost of the QM component of the calculation and simulations of explicit solvent molecules can hinder the efficiency of such calculations. The development of more efficient computational algorithm and computer hardware will enable fast and effective QM calculation, especially in cases where the size of the QM region is large. Also, finding

ways to automate this process will enable these calculations to be routinely applied to more ligands and enzyme targets—streamlining the process of drug design and development.

Bibliography

[1] Cohen, P. Protein kinases—the major drug targets of the twenty-first century? *Nat. Rev. Drug Discovery* **2002**, *1*, 309.

[2] Huang, J.; Lopes, P. E. M.; Roux, B.; Alexander D. MacKerell, J. Recent Advances in Polarizable Force Fields for Macromolecules: Microsecond Simulations of Proteins Using the Classical Drude Oscillator Model. *J. Phys. Chem. Lett.* **2014**, *5*, 3144–3150.

[3] Lemkul, J. A.; Huang, J.; Roux, B.; MacKerell Jr, A. D. An empirical polarizable force field based on the classical drude oscillator model: development history and recent applications. *Chem. Rev.* **2016**, *116*, 4983–5013.

[4] Alexov, E.; Mehler, E. L.; Baker, N.; M. Baptista, A.; Huang, Y.; Milletti, F.; Erik Nielsen, J.; Farrell, D.; Carstensen, T.; Olsson, M. H. M.; Shen, J. K.; Warwicker, J.; Williams, S.; Word, J. M. Progress in the Prediction of pKa Values in Proteins. *Proteins: Struct., Funct., Bioinf.* **2011**, *79*, 3260–3275.

Appendices

"One thing in life is for certain, the more profoundly baffled you have been in your life, the more open your mind becomes to new ideas."

— Neil deGrasse Tyson

A

CHARMM36 & AMBER99 Cysteine Topology

A.1 CHARMM36 Topology

A.1.1 CHARMM36 All-Hydrogen Cysteine Topology

```
RESI CYS          0.00
GROUP
ATOM N    NH1    -0.47  !     |
ATOM HN   H       0.31  !  HN-N
ATOM CA   CT1     0.07  !     |   HB1
ATOM HA   HB1     0.09  !     |   |
GROUP                   !  HA-CA--CB--SG
ATOM CB   CT2    -0.11  !     |   |     \
ATOM HB1  HA2     0.09  !     |   HB2    HG1
ATOM HB2  HA2     0.09  !   O=C
ATOM SG   S      -0.23  !     |
ATOM HG1  HS      0.16
GROUP
ATOM C    C       0.51
ATOM O    O      -0.51
BOND CB CA   SG CB   N HN  N  CA
BOND C  CA   C +N  CA HA  CB HB1
BOND CB HB2  SG HG1
DOUBLE O  C
IMPR N -C CA HN  C CA +N O
CMAP -C  N  CA  C   N  CA  C  +N
DONOR HN N
DONOR HG1 SG
ACCEPTOR O C
IC -C   CA   *N   HN    1.3479 123.9300  180.0000 114.7700  0.9982
IC -C   N    CA   C     1.3479 123.9300  180.0000 105.8900  1.5202
IC N    CA   C    +N    1.4533 105.8900  180.0000 118.3000  1.3498
IC +N   CA   *C   O     1.3498 118.3000  180.0000 120.5900  1.2306
IC CA   C    +N   +CA   1.5202 118.3000  180.0000 124.5000  1.4548
IC N    C    *CA  CB    1.4533 105.8900  121.7900 111.9800  1.5584
```

```
IC N    C    *CA  HA    1.4533 105.8900 -116.3400 107.7100  1.0837
IC N    CA   CB   SG    1.4533 111.5600  180.0000 113.8700  1.8359
IC SG   CA   *CB  HB1   1.8359 113.8700  119.9100 107.2400  1.1134
IC SG   CA   *CB  HB2   1.8359 113.8700 -125.3200 109.8200  1.1124
IC CA   CB   SG   HG1   1.5584 113.8700  176.9600  97.1500  1.3341
```

A.1.2 CHARMM36 Deprotonated Cysteine Topology

```
RESI CYSD        -1.00  ! Deprotonated Cysteine
(Thiolate charge modification based on ethylthiolate parameters, adm jr.)
GROUP
ATOM N    NH1    -0.47  !     |
ATOM HN   H       0.31  !  HN-N
ATOM CA   CT1     0.07  !     |   HB1
ATOM HA   HB1     0.09  !     |   |
GROUP                   !  HA-CA--CB--SG (-)
ATOM CB   CT2    -0.38  !     |   |     \
ATOM HB1  HA2     0.09  !     |   HB2    HG1
ATOM HB2  HA2     0.09  !   O=C
ATOM SG   S      -0.80  !     |
ATOM HG1  HS      0.00
GROUP
ATOM C    C       0.51
ATOM O    O      -0.51
BOND CB CA   SG CB   N HN  N  CA
BOND C  CA   C +N  CA HA  CB HB1
BOND CB HB2  SG HG1
DOUBLE O  C
IMPR N -C CA HN  C CA +N O
CMAP -C  N  CA  C   N  CA  C  +N
DONOR HN N
DONOR HG1 SG
ACCEPTOR O C
IC -C   CA   *N   HN    1.3479 123.9300  180.0000 114.7700  0.9982
```

```
IC -C   N    CA   C     1.3479 123.9300  180.0000 105.8900  1.5202
IC N    CA   C    +N    1.4533 105.8900  180.0000 118.3000  1.3498
IC +N   CA   *C   O     1.3498 118.3000  180.0000 120.5900  1.2306
IC CA   C    +N   +CA   1.5202 118.3000  180.0000 124.5000  1.4548
IC N    C    *CA  CB    1.4533 105.8900  121.7900 111.9800  1.5584
IC N    C    *CA  HA    1.4533 105.8900 -116.3400 107.7100  1.0837
IC N    CA   CB   SG    1.4533 111.5600  180.0000 113.8700  1.8359
IC SG   CA   *CB  HB1   1.8359 113.8700  119.9100 107.2400  1.1134
IC SG   CA   *CB  HB2   1.8359 113.8700 -125.3200 109.8200  1.1124
IC CA   CB   SG   HG1   1.5584 113.8700  176.9600  97.1500  1.3341
```

A.2 AMBER ff99SB-ILDNP Topology

A.2.1 AMBER ff99SB-ILDNP Cysteine Topology

```
[ CYS ]           0.00
[ atoms ]
N    N           -0.41570     1
H    H            0.27190     2
CA   CT           0.02130     3
HA   H1           0.11240     4
CB   CT          -0.12310     5
HB1  H1           0.11120     6
HB2  H1           0.11120     7
SG   SH          -0.31190     8
HG   HS           0.19330     9
C    C            0.59730     10
O    O           -0.56790     11
[ bonds ]
N      H
N     CA
CA    HA
CA    CB
CA     C
```

```
CB   HB1
CB   HB2
CB    SG
SG    HG
C      O
-C     N
[ impropers ]
-C    CA     N     H
CA    +N     C     O
```

A.2.2 AMBER ff99SB-ILDNP Deprotonated Cys Topology

```
[ CYM ]          -1.00    ! Deprotonated Cysteine
[ atoms ]
N    N           -0.41570     1
H    H            0.27190     2
CA   CT          -0.03510     3
HA   H1           0.05080     4
CB   CT          -0.24130     5
HB1  H1           0.11220     6
HB2  H1           0.11220     7
SG   SH          -0.88440     8
C    C            0.59730     9
O    O           -0.56790     10
[ bonds ]
N      H
N     CA
CA    HA
CA    CB
CA     C
CB   HB1
CB   HB2
CB    SG
 C     O
```

```
-C     N
[ impropers ]
-C    CA     N     H
CA    +N     C     O
```

A.3 Lennard-Jones Parameters for Cys Thiolate

Table A.1: Lennard-Jones parameters for selected atom types in cysteine thiolate

Force Field	Atom Name	Atom Type	Charge	ϵ (kJ mol^{-1})	σ (Å)
AMBER	CA	CT	-0.0351	0.4577	3.3997
	HA	H1	0.0508	0.0657	2.4714
	CB	CT	-0.2413	0.4577	3.3997
	SG	SH	-0.8844	1.0460	3.5636
CHARMM	CB	CS	-0.3800	0.4602	3.9200
	SG	SS	-0.8000	1.9665	3.9200

Table A.2: Charged residues within 5 Å of cysteine in the test set of PDB structures

Protein	Cys Residue	Charged Residues
α-1-AT	232	His231, Lys233, Lys234
ACBP-M46C	46	Glu41, Arg43, Glu48
ACBP-S65C	65	Lys66
ACBP-T17C	17	Lys18, Lys81
AhpC	46	Asp41, Glu49, Arg119, Arg142
HMCK	283	Glu232, Asp233
DJ-1	106	Glu18, His126, Arg156
Mb-G124C	124	Asp126
Mb-A125C	125	Asp126
MmsrA	72	Glu77, Glu115, Asp150
O^6-AGT	145	His146, Arg147, Lys165
Papain	25	Asp158, His159
ppΩ	25	Asp158, His159
PTP1B	215	Lys120, His214, Arg 221
YopH	403	His402, Arg404, Arg409

Table A.3: RMSD of protein backbone for the final coordinates of the protein structures from the RETI simulations of the thiol ($\lambda = 0$) and thiolate states ($\lambda = 1$).

Protein	**RMSD (Å)**	
	$\lambda = 0$	$\lambda = 1$
α-1-AT	1.51	1.51
ACBP-M46C	2.01	2.00
ACBP-S65C	2.10	1.93
ACBP-T17C	1.90	1.33
AhpC	4.22	4.27
DJ-1	1.42	1.73
HMCK	2.22	1.64
HMCK-S285A	2.26	1.74
Mb-G124C	2.12	2.02
Mb-A125C	2.22	1.97
MmsrA	4.11	4.90
MmsrA-E115Q	3.67	3.90
O^6-AGT	2.27	3.19
papain	0.79	0.72
ppΩ	1.34	0.84
PTP1B	1.15	1.01
YopH	0.98	1.72
YopH-H402A	1.41	1.37

Table A.4: Explicit solvent pK_a results with different histidine tautomeric states for tyrosine phosphatase proteins investigated.

Protein	**Cys Res.**	**Exptl. pK_a**	**His Residue**	**CHARMM**	**AMBER**
PTP1B	215	5.57 ± 0.12	H(S/I)D214	1.18 ± 0.40	1.31 ± 0.27
PTP1B	215	5.57 ± 0.12	H(S/I)E214	1.16 ± 0.77	4.14 ± 1.00
PTP1B	215	5.57 ± 0.12	H(S/I)P214	-1.70 ± 0.54	-0.76 ± 0.82
YopH	403	4.67 ± 0.15	H(S/I)D402	-1.16 ± 0.65	-1.78 ± 0.37
YopH	403	4.67 ± 0.15	H(S/I)E402	2.89 ± 0.71	4.63 ± 0.71
YopH	403	4.67 ± 0.15	H(S/I)P402	-2.80 ± 0.79	-5.29 ± 1.04
YopH-H402A	403	7.35 ± 0.04	H(S/I)D270,H(S/I)D350	-0.26 ± 0.59	1.69 ± 0.27
YopH-H402A	403	7.35 ± 0.04	H(S/I)E270,H(S/I)E350	0.25 ± 1.03	1.39 ± 0.39