November 2019

Constituent Partitioning Consensus Docking Models and Application in Drug Discovery

Rainer Metcalf
University of South Florida

Follow this and additional works at: https://scholarcommons.usf.edu/etd

Part of the Biochemistry Commons, and the Other Education Commons

Scholar Commons Citation

This Dissertation is brought to you for free and open access by the Graduate School at Scholar Commons. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Scholar Commons. For more information, please contact scholarcommons@usf.edu.
Constituent Partitioning Consensus Docking Models and Application in Drug Discovery

by

Rainer Metcalf

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Chemistry

Department of Chemistry
College of Arts and Sciences
University of South Florida

Co-Major Professor: Wayne Guida, Ph.D.
Co-Major Professor: Kenyon Daniel, Ph.D.
Yu Chen, Ph.D.
David Merkler, Ph.D.
Wesley Brooks, Ph.D.

Date of Approval:
November 14th, 2019

Keywords: Computational Chemistry Software, Virtual Molecular Modeling, Virtual Target Screening

Copyright © 2019, Rainer Metcalf
## Table of Contents

List of Tables .................................................................................................................................. vi

List of Figures .................................................................................................................................... viii

Abstract ........................................................................................................................................... xi

1 Introduction .................................................................................................................................. 1
  1.1 Preparation and Best Practices for Modeling .............................................................................. 2
    1.1.1 Protein Structure Selection ................................................................................................. 2
    1.1.2 X-Ray Crystal Structure Resolution .................................................................................. 3
    1.1.3 Co-Factors, Ligand and Sequence Verification, and Biological Assembly ......................... 4
    1.1.4 Model Optimization ........................................................................................................... 5
    1.1.5 Visual Interfaces and User Environments .......................................................................... 6
  1.2 Selection of Molecular Mechanics Force Field ............................................................................. 8
    1.2.1 Force Field Energy Components ........................................................................................ 9
    1.2.2 Generalized Force Fields ..................................................................................................... 10
    1.2.3 Customized Parameterization of Force Fields .................................................................... 12
    1.2.4 More Complex Models ........................................................................................................ 14
  1.3 Modeling Methods, Algorithms, and Techniques ......................................................................... 16
    1.3.1 Protein and Ligand Preparation .......................................................................................... 16
      1.3.1.1 Determination of Protonation States ........................................................................... 17
    1.3.2 Energy Minimization ........................................................................................................... 17
      1.3.2.1 Energy, Entropy, Free Energy, and Enthalpy ............................................................... 18
      1.3.2.2 Energy Minimization Algorithms ................................................................................ 20
      1.3.2.3 MM/GBSA .................................................................................................................... 22
      1.3.2.4 Low Frequency Normal Mode Analysis ....................................................................... 23
    1.3.3 Molecular Dynamics and Monte Carlo .................................................................................. 25
      1.3.3.1 MD vs MC ..................................................................................................................... 25
      1.3.3.2 Simulated Annealing ..................................................................................................... 27
      1.3.3.3 Equilibration ................................................................................................................ 28
      1.3.3.4 Common Setup Parameters for MD Simulation ........................................................... 29
      1.3.3.5 Trajectory Analysis for Equilibration .......................................................................... 33
    1.3.4 Quantum Chemistry ............................................................................................................ 35
      1.3.4.1 Quantum Calculation Methods .................................................................................... 36
      1.3.4.2 Basis Sets ..................................................................................................................... 39
    1.3.5 Quantum Mechanical Molecular Modeling (QM/MM) ......................................................... 40
    1.3.6 Ligand Docking .................................................................................................................... 43
      1.3.6.1 SP and XP Docking ....................................................................................................... 48
1.3.6.2 Induced Fit Docking ................................................................. 49
1.3.6.3 Quantum Polarized Ligand Docking .................................... 50
1.3.6.4 Constituent Partitioning Consensus Docking ....................... 51
1.3.7 Virtual Target Screening ........................................................... 55
  1.3.7.1 Calibration ........................................................................ 55
  1.3.7.2 Scoring and Confidence Intervals ...................................... 59
  1.3.7.3 Ligand Specificity .............................................................. 66
  1.3.7.4 Enzyme Selectivity ............................................................ 67
1.3.8 Homology Modeling and Protein Folding ................................ 68
1.4 Informatics .................................................................................. 73
  1.4.1 Chemoinformatics ................................................................. 74
    1.4.1.1 Pharmacophore Modeling .............................................. 75
  1.4.2 Bioinformatics ...................................................................... 76
  1.4.3 Best Practices in Statistics ..................................................... 78
1.5 Summary ..................................................................................... 79
1.6 Acknowledgments ....................................................................... 81

2 The Atypical Protein Kinase C Small Molecule Inhibitor Z-Stat, and Its Effects on Invasion Through Decreases in PKC-ζ Protein Expression ................................................. 82
  2.1 Introduction ............................................................................. 82
  2.2 Materials and Methods ............................................................. 85
    2.2.1 Antibodies and Reagents .................................................... 85
    2.2.2 Analysis of Somatic Gene Mutations for Ovarian Cancer, CCOC and CCOC Cell Lines ................................................................. 86
    2.2.3 Computational Analysis of aPKCs ......................................... 86
      2.2.3.1 Protein Preparation ...................................................... 86
      2.2.3.2 Molecular Dynamics .................................................... 87
      2.2.3.3 Consensus Docking ..................................................... 88
      2.2.3.4 Rigid Receptor Docking (RRD) ................................... 89
      2.2.3.5 Induced Fit Docking (IFD) .......................................... 89
      2.2.3.6 Quantum Polarized Ligand Docking (QPLD) .................. 90
      2.2.3.7 Molecular Mechanics and Generalized Born Surface Area (MM/GBSA) ................................................................. 91
      2.2.3.8 Virtual Target Screening (VTS) ................................... 92
    2.2.4 Cell Culture ....................................................................... 92
    2.2.5 Atypical PKC Expression During Rapid Growth and Cell Cycle Arrested ................................................................. 93
    2.2.6 Preliminary Screening of TOV21G and ES-2 with ICA-1S and Z-Stat ................................................................. 93
    2.2.7 Cell Viability Assay ............................................................ 93
    2.2.8 Cell Lysate Collection ........................................................ 93
    2.2.9 Western Blot Analysis ....................................................... 94
    2.2.10 Endpoint PCR ................................................................. 94
    2.2.11 Rac1 Activation Assay ...................................................... 95
    2.2.12 Preparation of cytoplasmic and nuclear extracts .................. 95
    2.2.13 Fluorescent Microscopy .................................................... 95
2.2.14 Invasion Assay .................................................................96
2.2.15 ζ-Stat in-vivo .................................................................96
2.2.16 Statistical Analysis ..........................................................97
2.3 Results .................................................................................98
  2.3.1 PIK3CA and ARID1A Are in the Top Mutated Genes in All Ovarian Tissue Types and in CCOC ..........................98
  2.3.2 Select Inhibitor Effects on PKC-ζ and PKC-ι Protein Expression .................................................100
  2.3.3 In-silico Results and Model Validation .................................................................101
    2.3.3.1 Model Viability .................................................................101
    2.3.3.2 Docking Results .................................................................108
    2.3.3.3 Virtual Target Screening Results .................................................................110
  2.3.4 Inhibition of Cell Viability ......................................................111
  2.3.5 Analysis of the PKC-ζ/Ect2/Rac1/RhoA Pathway .................................................................112
  2.3.6 Analysis of ζ-Stat in TOV21G Tumor Xenografts .................................................................116
  2.3.7 ζ-Stat Interrupts the PKC-ζ/Ect2 via PKC-ζ Protein Decrease .........................................................118
2.4 Discussion ..........................................................................119
  2.4.1 Targeting aPKCs in CCOC .................................................................119
  2.4.2 aPKCs are Involved in the Localization of Ect2 .................................................................120
  2.4.3 ζ-Stat has Therapeutic Potential in CCOC .................................................................121
2.5 Acknowledgments .................................................................123
  2.5.1 Funding .................................................................123
  2.5.2 A Note to Reader .................................................................123
2.6 Availability of Data and Materials ......................................................123
2.7 Conflict of Interest ................................................................123
2.8 Author Contributions .................................................................124
2.9 Supplemental Figures .................................................................125

3 Ligand-Mediated Protein Degradation Reveals Functional Conservation Among Sequence Variants of the CUL4-type E3 Ligase Substrate Receptor Cereblon ................................129
  3.1 Introduction ................................................................129
  3.2 Results ...........................................................................131
    3.2.1 Resistance of Mouse Cells to Immunomodulatory Compounds ................................131
    3.2.2 Thalidomide-Binding Domain of CRBN has a Conserved Immunomodulatory Compound–Binding Motif .................................................................135
    3.2.3 Immunomodulatory Compound Binding is Conserved in CRBN Sequence Variants .................................................................136
    3.2.4 N-terminal Stabilization of CRBN–Immunomodulatory Compound Interactions .................................................................140
    3.2.5 Acquired Ubiquitin-Proximity Ligation in Mouse Cells Establishes Conserved Ligase Functions of Mouse and Human CRBN Variants .................................................................142
  3.3 Discussion ........................................................................147
  3.4 Experimental Procedures .................................................................150
    3.4.1 Animals and Cell Lines .................................................................150
    3.4.2 T-Cell Isolation, Activation, and Drug Treatments .................................................................151
    3.4.3 Quantitative Real-Time PCR .................................................................152
    3.4.4 Treatment of Multiple Myeloma Cells .................................................................152
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.4.5 General Chemistry Information</td>
<td>153</td>
</tr>
<tr>
<td>3.4.6 Cloning, Protein Expression, And Purification</td>
<td>154</td>
</tr>
<tr>
<td>3.4.7 Zincon assay</td>
<td>155</td>
</tr>
<tr>
<td>3.4.8 ITC</td>
<td>156</td>
</tr>
<tr>
<td>3.4.9 Intrinsic Tryptophan Fluorescence Assay</td>
<td>156</td>
</tr>
<tr>
<td>3.4.10 Molecular Modeling Preparation</td>
<td>157</td>
</tr>
<tr>
<td>3.5 Author Contributions</td>
<td>158</td>
</tr>
<tr>
<td>3.6 Acknowledgments</td>
<td>159</td>
</tr>
<tr>
<td>3.7 Supplementary Materials</td>
<td>160</td>
</tr>
<tr>
<td>3.7.1 Supplementary Methods</td>
<td>160</td>
</tr>
<tr>
<td>3.7.1.1 Protein Preparation</td>
<td>160</td>
</tr>
<tr>
<td>3.7.1.2 Molecular Dynamics</td>
<td>161</td>
</tr>
<tr>
<td>3.7.1.3 Consensus Docking</td>
<td>163</td>
</tr>
<tr>
<td>3.7.1.4 Induced Fit Docking</td>
<td>164</td>
</tr>
<tr>
<td>3.7.1.5 Rigid Receptor Docking</td>
<td>165</td>
</tr>
<tr>
<td>3.7.1.6 Quantum Polarized Ligand Docking</td>
<td>165</td>
</tr>
<tr>
<td>3.7.1.7 Molecular Mechanics and Generalized Born Surface Area</td>
<td>166</td>
</tr>
<tr>
<td>3.7.1.8 Detailed Chemical Methods</td>
<td>167</td>
</tr>
<tr>
<td>3.7.1.9 Detailed Protein Production of Wild-Type and Mutant CRBN</td>
<td>170</td>
</tr>
<tr>
<td>3.7.2 Supplementary Figures</td>
<td>171</td>
</tr>
<tr>
<td>3.8 Chapter Appendix</td>
<td>184</td>
</tr>
<tr>
<td>3.8.1 Supplementary Computational Results</td>
<td>184</td>
</tr>
<tr>
<td>3.8.2 Determination of Equilibration for MD Systems</td>
<td>184</td>
</tr>
<tr>
<td>3.8.2.1 Equilibration of hCRBN Model System</td>
<td>184</td>
</tr>
<tr>
<td>3.8.2.2 Equilibration of mCRBN Model System</td>
<td>188</td>
</tr>
<tr>
<td>3.8.2.3 Equilibration of gCRBN Model System</td>
<td>189</td>
</tr>
<tr>
<td>3.8.2.4 Equilibration of hmCRBN Model System</td>
<td>190</td>
</tr>
<tr>
<td>3.8.2.5 Equilibration of haCRBN model system</td>
<td>190</td>
</tr>
<tr>
<td>3.8.3 Determination of H378 Protonation States</td>
<td>192</td>
</tr>
<tr>
<td>3.8.4 Computational Model Validation</td>
<td>194</td>
</tr>
<tr>
<td>3.8.5 CP Consensus Docking Results</td>
<td>195</td>
</tr>
<tr>
<td>3.8.6 High Accuracy Course-Grained Docking Derived from CP Consensus Docking</td>
<td>197</td>
</tr>
<tr>
<td>3.8.7 Conclusions from Modeling</td>
<td>198</td>
</tr>
</tbody>
</table>

4 Discovery of a Non-Nucleotide Potential STING Antagonist from Consensus Docking and Dimer Identification with Novel Site-Restriction Virtual Screening 200

4.1 Introduction 200

4.2 Experimental Design and Methods 207

4.2.1 Computational Methods 207

4.2.1.1 Protein Preparation 207

4.2.1.2 Molecular Dynamics 208

4.2.1.3 Consensus Docking 209

4.2.1.4 Rigid Receptor Docking 209

4.2.1.5 Induced Fit Docking 210
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.2.1.6 Quantum Polarized Ligand Docking</td>
<td>211</td>
</tr>
<tr>
<td>4.2.1.7 Molecular Mechanics and Generalized Born Surface Area</td>
<td>211</td>
</tr>
<tr>
<td>4.2.2 Experimental Methods</td>
<td>212</td>
</tr>
<tr>
<td>4.2.2.1 Surface Plasmon Resonance</td>
<td>212</td>
</tr>
<tr>
<td>4.2.2.2 Microscale Thermophoresis</td>
<td>213</td>
</tr>
<tr>
<td>4.2.2.3 Luciferase Assay</td>
<td>213</td>
</tr>
<tr>
<td>4.2.2.4 Chemical Synthesis of NSC335504</td>
<td>214</td>
</tr>
<tr>
<td>4.3 Results &amp; Discussion</td>
<td>215</td>
</tr>
<tr>
<td>4.3.1 Computational Model Construction and Validation</td>
<td>215</td>
</tr>
<tr>
<td>4.3.2 STING Isoform Selection</td>
<td>217</td>
</tr>
<tr>
<td>4.3.3 Consensus Docking Results</td>
<td>221</td>
</tr>
<tr>
<td>4.3.4 Site-Restriction Virtual Screening for Identification of Possible Dimer Complexes</td>
<td>224</td>
</tr>
<tr>
<td>4.3.5 Direct Binding Assays</td>
<td>226</td>
</tr>
<tr>
<td>4.3.6 THP-1 Luciferase Reporter Assay</td>
<td>228</td>
</tr>
<tr>
<td>4.4 Conclusions</td>
<td>229</td>
</tr>
<tr>
<td>4.5 Supplementary Material</td>
<td>231</td>
</tr>
<tr>
<td>4.5.1 Supplemental Figures</td>
<td>231</td>
</tr>
</tbody>
</table>

References | 234 |

Appendices | 277 |

Appendix 1: Mixing Properties of Sphingomyelin Ceramide Bilayers: A Simulation Study | 277 |

Appendix 2: Evaluation of digestion methods for analysis of trace metals in mammalian tissues and NIST 1577c | 288 |

Appendix 3: Cupriphilic compounds to aid in proteasome inhibition | 295 |

Appendix 4: Fair Use Worksheets | 300 |

Appendix 4.1: Ligand-Mediated Protein Degradation Reveals Functional Conservation Among Sequence Variants of the CUL4-type E3 Ligase Substrate Receptor Cereblon | 300 |

Appendix 4.2: Mixing Properties of Sphingomyelin Ceramide Bilayers: A Simulation Study | 303 |

Appendix 4.3: Evaluation of digestion methods for analysis of trace metals in mammalian tissues and NIST 1577c | 306 |

Appendix 4.4: Cupriphilic compounds to aid in proteasome inhibition | 309 |
List of Tables

Table 1: Software that implements particular force fields .................................................................14
Table 2: Table of available MD Software ............................................................................................27
Table 3: Overview of molecular timescales for general atomic and protein movements ...........30
Table 4: List of quantum chemistry software ....................................................................................40
Table 5: List of docking software .......................................................................................................45
Table 6: Interpretations of Binding Site Characteristics from CP Consensus Docking .................52
Table 7: Software for computational modeling ..................................................................................81
Table 8: Six most common gene mutations in all ovarian cancers ...............................................99
Table 9: Somatic mutations in CCOC and CCOC cell lines ...............................................................100
Table 10: Binding affinity of immunomodulatory compounds to WT and mutant human CRBN-TBD determined using ITC and fluorescence intensity assay .........................139
Table 11: IC$_{50}$ of Multiple Myeloma Cell Lines ............................................................................173
Table 12: Pose RMSD Calculations and Docking Scores ................................................................177
Table 13: Statistical analysis of human T cells treated with IMiDs, JQ1, and dBET .................182
Table 14: Statistical analysis of CRBN$^{+/+}$ and CRBN$^{-/-}$ mouse T cells treated with IMiDs, JQ1, and dBET ..............................................................................................................183
Table 15: CP Consensus Docking on hCRBN with controls and experimental K$_D$ values ........196
Table 16: Course-grained virtual screening model utilizing only XP docking for assessing analog binding affinity .......................................................................................................................199
Table 17: List of PDB entries used for MD simulation .......................................................................207
Table 18: Model comparisons to literature and experiment ...............................................................220
Table 19: Consensus docking using disparate docking methodologies to partition physical characteristics of the STING-CDN binding site .........................................................222
List of Figures

Figure 1: Diversity analysis of SMU and NCI diversity set ..........................................................58
Figure 2: Gaussian distribution fit of calibration set average docking scores ............................62
Figure 3: Log-normal distribution fit of calibration set standard deviations .............................64
Figure 4: Power law distribution fit of self dock RMSD ...............................................................65
Figure 5: Representative normal distribution fits of calibration set docking energies ...........69
Figure 6: Atypical PKC inhibitors ...............................................................................................84
Figure 7: PKC-ζ and PKC-ι protein expression in rapidly growing and serum starved cells and the effects of ICA-1S and ζ-Stat on PKC-ζ and PKC-ι protein expression in SHT290, TOV21G and ES-2 CCOC cell lines ........................................................101
Figure 8: RMSD of backbone atoms and Radius of Gyration of all atoms graphs for PKCι Molecular Dynamics simulation .........................................................................................102
Figure 9: RMSF (Root Mean Square Fluctuation) plot for PKCι Molecular Dynamics simulation ..................................................................................................................103
Figure 10: RMSD of backbone atoms and Radius of Gyration of all atoms graphs for PKCζ Molecular Dynamics simulation .........................................................................................105
Figure 11: RMSF (Root Mean Square Fluctuation) plot for PKCζ Molecular Dynamics simulation ..................................................................................................................106
Figure 12: Computational modeling results ................................................................................107
Figure 13: Cell proliferation and viability of SHT290, TOV21G and ES-2 CCOC cell lines .................................................................................................................................112
Figure 14: Effects of ζ-Stat on Ect2 localization and RhoA protein and genomic expression .................................................................................................................................114
Figure 15: Immunofluorescent staining of Ect2 ........................................................................115
Figure 16: PKC-ζ regulates invasion and migration of TOV21G ovarian cancer ....................116
Figure 17: Effects of ζ-Stat on in-vivo xenografts with athymic nude mouse models ..........117

Figure 18: Hypothesized pathway illustrating the scaffolding of Ect2 via PKC-ζ in the cytoplasm .................................................................118

Figure 19: Raw Western blot images for TOV21G and SHT290 Figure 7A ...................125

Figure 20: Raw Western blot images for ES-2 Figure 7A ........................................126

Figure 21: Raw Western blot images for TOV21G Figure 7B ..................................126

Figure 22: Raw Western blot images for ES-2 Figure 7B ........................................127

Figure 23: Raw Western blot images for TOV21G Rac1 PD utilizing GST PAK-PBD Figure 16D .................................................................127

Figure 24: Raw Western blot images for TOV21G RhoA and PKC-ζ protein decrease ....128

Figure 25: Mouse cells are resistant to immunomodulatory drugs ............................133

Figure 26: Structure of cereblon is conserved across different species ....................134

Figure 27: Human and mouse CRBN binds IMiDs with similar affinities ...............138

Figure 28: Lenalidomide binds to the TBD and CRBN–DDB1 protein complex ..........141

Figure 29: Functional activity of lenalidomide, JQ1, and dBET1 in human T cells and binding affinities of JQ1, dBET1, and Len to human CRBN ...............................143

Figure 30: Functional activity of lenalidomide, pomalidomide, JQ1, and dBET1 in mouse T cells ........................................................................146

Figure 31: Synthesis of tert-Butyl (2-(2,6-dioxopiperidin-3-yl)-1-oxoisindolin-4-yl) carbamate ........................................................................167

Figure 32: Structure of Me-dBET1 ...........................................................................169

Figure 33: Sequence conservation of CRBN thalidomide binding site .....................171

Figure 34: Dose of anti-CD3 and lenalidomide to induce IL-2 and proliferation by T cells .......................................................................................173

Figure 35: RMSD and Rg profiles of hCRBN, mCRBN, gCRBN, and hmCRBN ..........174

Figure 36: RMSF profiles of hCRBN, mCRBN, gCRBN, and hmCRBN .....................175
Figure 37: Potential energy profiles of hCRBN, mCRBN, gCRBN, and hmCRBN ..............175
Figure 38: Hydrogen bonding analysis of Len isoindolinone amine and E377 carboxyl
group ..................................................................................................................................176
Figure 39: Recombinant TBD protein expression and purification and zinc binding
domain analysis using the zinccon assay ...........................................................................178
Figure 40: Binding affinity of IMiDs to TBD variants by ITC ........................................179
Figure 41: Binding pocket with N-methyl-lenalidomide ...................................................180
Figure 42: Drug Treatment of human and mouse T cells and multiple myeloma cell lines ......181
Figure 43: 3D chemical structures of IMiDs and CRBN ribbon overlays ..........................186
Figure 44: Overlays with centered view of the CRBN thalidomide binding site ..........187
Figure 45: SiteMap binding site characterization and overlays of human and mouse
CRBN ....................................................................................................................................188
Figure 46: Equilibration profiles for haCRBN model system .........................................192
Figure 47: STING pathway .................................................................................................203
Figure 48: SiteMap analysis between CDN binding sites of hSTING and mSTING ..........206
Figure 49: Clonixeril synthesis scheme ..............................................................................215
Figure 50: RMSF profiles between REF, WT, and HAQ isoforms of STING bound to
2’3’-cGAMP .......................................................................................................................218
Figure 51: hSTING^REF CTD structures after MD ..........................................................219
Figure 52: Site-restriction docking method for identifying potential dimeric ligand
complexes ..............................................................................................................................225
Figure 53: Molecular representation of NSC335504 and associated mass spectrum
profile ....................................................................................................................................226
Figure 54: MST and SPR binding assays for NSC335504 .................................................227
Figure 55: Steady state SPR binding affinity plots for c-di-GMP and 2’3’-cGAMP
controls .................................................................................................................................227
Figure 56: 2′3′-cGAMP and NSC335504 induced IRF3 upregulation in THP-1 cells ..........229

Figure 57: RMSD and Radius of Gyration plots for MD systems ..............................................231

Figure 58: H185 Cα distances for MD systems ........................................................................232

Figure 59: H185 Cα distance for STINGREF apo structure MD system .................................233
Abstract

This work expounds on some of the current computational tools and programs available and the best practices associated with their use. A high-level introduction, intended for both novices and the semi-experienced, focusing on the more common programs used in scientific literature is the scope of this topic. Both classical and quantum techniques are described. Classical methodologies include Molecular Dynamics, Monte Carlo, energy minimization methods, molecular docking, low-mode, and homology modeling. Quantum chemistry techniques are also discussed encompassing Hartree-Fock, Post-Hartree-Fock theories, and Density Functional Theory along with associated basis sets.

Along with established methodologies, novel theoretical methods are introduced for furthering the application of computational modeling. Constituent partitioning consensus docking makes use of disparate docking methodologies to elucidate physical characteristics of protein binding sites. This opus also advances the function of virtual target screening, implementing robust algorithmic treatment of the protocol and improving the accuracy and scope of target identification and binding site description. The introduction focuses on theoretical approaches while subsequent chapters encompass the execution of these techniques in practical applications of drug discovery.
1 Introduction

This chapter expounds on some of the current computational tools and programs available and the best practices associated with their use. A high-level introduction focusing on the more common programs used in scientific literature is the scope of this topic. Additionally, the chapter concentrates on biochemistry and protein dynamics and does not include references to solid state modeling.

Proteins might be best characterized on the molecular scale as incredible, awe-inspiring, and deceptively simple wriggling balls of snot and frustration. Many rarely exist in conformations that allow catalysis (1–6), yet all life depends on their function which, in turn, depends almost entirely on a stochastic 3D morphology dictated purely by quantum effects leading to minute energetic biases toward an overall statistically ephemeral structure. As such, exploring the properties of these entities is an insane task.

Yet, in the face of such lunacy, we attempt the equally absurd approach of molecular modeling that makes use of all theoretical methods and computational techniques to model or mimic the behavior of these infinitely complex molecules. It studies molecular systems ranging from small chemical systems to large biological molecules requiring the aid of computers to process the vast quantities of data and intricate algorithms. While the assumption that our reality can be readily explained by microscale deterministic events is a reasonable one, modeling these events and extrapolating their outcomes to describe bulk properties and predict macroscale quantities is an arduous undertaking with consistently meager results. From experience, much of
the failings of computational models appears to be from thinking only one program or algorithm is needed for any particular task at hand. Just as those large-scale properties we wish to elucidate are not as conveniently simple as most might like, nor too is the modeling and practices needed to surmount such Brobdingnagian tasks.

1.1 Preparation and Best Practices for Modeling

First and foremost, when developing a computational model, preparation is key. Molecular modeling is highly dependent on experimental data for initial structure. Therefore, proper identification of the specific protein and/or ligand is crucial before moving forward. As always, this necessitates an extensive scientific literature search including the complete name (along with all know aliases), sequence (full biological assembly and individual domains), and origins of protein/ligand (species, tissue, cellular compartment, metabolic pathway), binding sites (orthosteric, known allosteric, and any potential alternate sites). For ligands, chemical structure, chirality, tautomerization, and protonation states of the ligand should be accounted for. The more information, the better.

1.1.1 Protein Structure Selection

For protein structures, an exhaustive wwPDB (Worldwide Protein Data Bank (7), portal sites through RCSB (8), ePDB, and jPDB) search for the specified protein under all known aliases, all possible isozymes, and/or interspecies homologs is required. If no structure exists, molecular modeling is severely limited (see section 1.3.8 Homology Modeling and Protein Folding for more details). X-ray diffraction studies are preferable for a starting point due to finer structural detail and better approximation of resolution. NMR (Nuclear Magnetic Resonance) is
generally more useful as an addendum to X-ray diffraction by providing information about protein loop and side chain motion (9). Additionally, NMR is currently the only way to study intrinsically disordered proteins, which are estimated to comprise 40–50% of the human proteome (10,11). Lately, cryo-electron microscopy (cryo-EM) has made significant technological leaps and will likely replace X-ray crystallography as the standard for structure determination in the near future. However, at the time of this publication, most cryo-EM structures do not have comparable resolution to X-ray diffraction and are only amenable to studies of mesoscopic scales. This technique is indispensable for understanding protein-protein interactions and complexes, but currently not able to replace a fine-tuned crystallographic study (12,13).

1.1.2 X-Ray Crystal Structure Resolution

Optimal crystallographic resolution of a structure for modeling purposes is less than 2Å. For structural resolution between 2Å and 3Å, significant uncertainties exist in the crystal structure and further computational refinement and control studies are necessary (outlined later in this chapter). Without proper refinement and controls, substantial error will likely persist throughout a modeling study because of possible inaccuracies in initial protein structure, such as indeterminate protonation states and disordered loop conformations (14). Protonation states are designated by inference any way since hydrogen atoms are not localized even at these resolutions. For structural resolutions beyond approximately 3.5Å, errors in the crystal structure tend to be too large for proper computational refinement. At this point, electron density is far too poor for localizing even internal residues leading to domain regions becoming unstable and any information on loop structures is typically completely lost (15).
1.1.3 Co-Factors, Ligand and Sequence Verification, and Biological Assembly

Many other factors aside from structural resolution must be examined before any modeling or experimentation is performed, e.g. verification of sequence (including allotypes, isozymes and possible mutants) and ligand binding (holo and/or apo structures) of target protein.

If a PDB entry is missing less than 2-3 residues owing to a lack of structural resolution or intended methodological exclusion, homology modeling software packages, such as Schrödinger's Prime (16) and YASARA (17,18) or the web-based SWISS-MODEL (19,20), can be used to estimate possible initial configurations of known sequences. If any of the missing residues comprise part of the active site or exceed 2-3 missing residues, quantum mechanical modeling and refinement (Sections 1.3.4 Quantum Chemistry, 1.3.5 Quantum Mechanical Molecular Modeling, and 1.3.8 Homology Modeling and Protein Folding) may be necessary. Another common oversight is a failure to identify potential crystal artifacts that may unnaturally influence the crystallographic structure. Crystal packing can cause unphysical dimerization or induced folding not seen in nature. There may also exist possible dependencies on dimerization for protein function. An extensive review of literature for understanding the biological environment, pathways, and potential binding partners is vital to determine the biological assembly of a target protein. Protein complexes can form alternate folding or binding pockets relevant for study, but their addition will obviously complicate any modeling. Great care is needed for optimization of computational costs versus model detail.
1.1.4 Model Optimization

For computation, there is always a compromise between accuracy and speed/cost of calculation. Minimizing the size and complexity of a system, using the fewest calculations possible, and specializing a model to a specific task with the greatest number of assumptions, all while maintaining a sufficient level of accuracy is a discussion of high priority. Three general categories arise when discussing optimization. A simulation or model can be completed quickly, cheaply (in terms of hardware or financial resources), or accurately. Realistically, prioritization can only be done on two of three categories at best.

Speed is largely determined by size and complexity of a system and the dynamics to be studied. Algorithm selection and level of applied theory is helpful here. For example, Molecular Dynamics (MD) simulation is faster than quantum mechanical simulation. System simplification or focusing on only the most pertinent kinetics can also reduce simulation time so long as other assumptions are reasonably held. Cost is generally the simplest way to solve an optimization problem. A bigger computer with better hardware is an effective solution for most research. Faster algorithms can also be implemented, but many require significant manpower and costly research to optimize and some software applications under a commercial license are simply more costly than others. Accuracy is the most difficult of the three to address. This is where proper statistical methodology can help optimize sample size or extrapolate a data set (see section 1.4.3 Best Practices in Statistics for details). When it comes to determination of physical quantities, however, universal constraints and uncertainty limit precise determination of any physical measurements. For thermodynamic quantities, such as binding affinity, the trade-off between precision and dissipation is the thermodynamic uncertainty relation. This relation shows that a more precise output requires a higher thermodynamic cost independent of the time used to
produce the output (21). Therefore, no experiment can have a true free energy resolution of an affinity lower than $k_B T$. For structural determination, low electron density around hydrogen atoms restrict hydrogen atom resolution as well as the need for extremely high energy X-rays to reach resolutions lower than 1Å.

It should also be noted that crystal structures can be notoriously inaccurate as a result of the extreme conditions necessary to crystallize a protein that are far from biologically relevant (e.g. low temperatures, high/low pH, large salt concentrations, crystal artifacts caused by protein-protein interactions of crystal mates, the fact that solid crystals do not exist in living cells, etc.) (22). It is for these reasons that crystallography is also highly dependent on computational modeling for verifying aspects of crystallographic data. System equilibration should always be standard under MD simulation before further modeling (see section 1.3.3.3 Equilibration). This mutual dependence on theory and experimentation requires close collaboration with both areas of study.

1.1.5 Visual Interfaces and User Environments

Programs that are primarily for 3D viewing are considered visual interfaces and user environment programs provide a platform for interfacing with other programs. These are not considered molecular modeling software in and of themselves, simply a way to view or access other modeling software output.

The molecular visualization system, PyMOL (23), is open-source software released under the Python license. The program was initially developed by DeLano Scientific LLC, a private software company dedicated to making accessible tools specifically for scientific and educational uses, and is currently commercialized by Schrödinger, Inc. PyMOL can produce high-quality 3D
images of small molecules and biological macromolecules and is able to solve Poisson–Boltzmann equations for electrostatics visualization. However, the program is not intended for molecular modeling.

VMD (Visual Molecular Dynamics) is another popular molecular modeling and visualization computer program used primarily as an interface for many third-party, open source molecular modeling applications (24). VMD was first developed as a tool to view and analyze the results of molecular dynamics simulations. More recently, it now includes an extensive tool set for working with more rigorous applications such as volumetric data, sequence data, and arbitrary graphics objects. Development of several external plug-ins and an increasing set of features and tools make it one of the most commonly used software packages in computational chemistry, biology, and biochemistry.

In addition to PyMOL, Schrödinger, Inc. also develops another graphical user interface called Maestro (25), but it is more akin to VMD as a single common user environment providing a platform for fully-integrated molecular modeling, visualization, and analysis. Another similar and popular proprietary software that serves as both a visualization program and platform for a range of modeling and simulation applications is MOE (Molecular Operating Environment) (26).

Another common interface for computational chemistry is Avagadro (27), a molecular builder and editor based on Open Babel. It is designed for cross-platform use in computational chemistry, molecular modeling, bioinformatics, materials science, and related areas and utilizes a plug-in architecture to extend its capabilities.
1.2 Selection of Molecular Mechanics Force Fields

The force terms and integrated classical constraints, such as bonding interactions, bond angles, vibrational spring constants, dihedral angles, partial charges, etc., make up a force field file for specific molecules. The force field files are built from a set of potential energy equations and their derivatives combined with the parameters used to describe pairwise interactions between unique particles or atoms. There are essentially three main types of force fields with varying atomic resolution. All atom force fields are the most common, where each atom is enumerated with a parameter set. United atom force fields provide parameters for all atoms except non-polar hydrogens, which are folded into the parameter set for the larger bonded atom creating a more coarse grained model. Lastly, there are fully coarse grained parameter sets where several atoms are grouped together as an abstract representation of a single particle.

The parameters related to these force fields rely heavily on the accuracy and availability of experimental data for individual molecules and bulk effects. Tuning parameters is usually semi-empirical, application-specific, and is where differences between force fields are most apparent. For example, OPLS was originally tuned to reproduce certain bulk properties of liquids, whereas CHARMM was tuned mainly for macro-molecules, such as proteins and nucleic acids. Therefore, it can be argued that certain force fields may be better or worse for specific applications. OPLS might be more suited for liquid studies, CHARMM for proteins, and AMBER for DNA simulations. However, even with these disparate approaches, no substantial difference has been observed between the force fields like AMBER, CHARMM, OPLS, GROMOS, etc., used in computational biophysics (28). The more immediate and substantial sources of error come from model design, implementation, and data interpretation.
1.2.1 Force Field Energy Components

The main components of a force field are bond distances and bond angles, described by harmonic spring potentials due to near ideal stretching behavior and slight fluctuations around equilibrium values observed at temperatures generally used for molecular simulations. These parameters can be constrained to fixed values to save computational effort with some loss of accuracy. A so-called improper dihedral term accounts for out of plane bending to help keep certain atom groups, such as aromatic rings and amides, planar. The dihedral potential is characterized by a cosine expansion whose amplitude describes the barrier height between the low energy conformations. Long range interactions specifically enumerated in the force field file are only counted for atoms three or more bonds apart, consisting of Coulomb and Lennard-Jones two-body interaction terms. The Lennard-Jones potential is a combination of attractive van der Waals forces and empirical repulsive forces due to Pauli repulsion.

General Classical MM Hamiltonian

\[ E_{\text{Total}} = E_{\text{Bond}} + E_{\text{Angle}} + E_{\text{Imp\,Dih}} + E_{\text{Dihedral}} + E_{\text{Coulomb}} + E_{\text{VdW}} \]  

(Eq. 1)

Where,

\[ E_{\text{Bond}} = \sum_{i<j} K_{\text{Bond}_{ij}} \left( r_{ij} - r_0 \right)^2 \]  

(Eq. 2)

\[ E_{\text{Angle}} = \sum_{i<j<k} K_{\theta_{ijk}} \left( \theta_{ijk} - \theta_0 \right)^2 \]  

(Eq. 3)

\[ E_{\text{Imp\,Dih}} = \sum_{ijkl} K_{\zeta_{ijkl}} \left( \zeta_{ijkl} - \zeta_0 \right)^2 \]  

(Eq. 4)
\[ E_{\text{Dihedral}} = \sum_{n=0}^{N} K_n \cos(\omega)^n \] (Eq. 5)

\[ \approx \sum_{ijkl} \frac{1}{2} [K_1 (1 + \cos(\omega)) + K_2 (1 - \cos(2\omega)) + K_3 (1 + \cos(3\omega)) + K_4] \]  

\[ E_{\text{Coulomb}} = \sum_{i<j} \frac{q_i q_j}{4\pi \epsilon_o r_{ij}} \] (Eq. 6)

\[ E_{\text{vdW}} = \sum_{i<j} 4\epsilon_o \left[ \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \] (Eq. 7)

The OPLS potential function (and some formulations of AMBER) is given as the first few terms of a Fourier series of the Ryckaert-Bellemans function. OPLS parameters were originally the exact AMBER parameters for bonded interactions, but the two formulations have slowly diverged over the years. OPLS was originally known as OPLSA (A meaning AMBER). An alternate dihedral formulation is used for some force fields (CHARMM in particular) and is ultimately equivalent but has the advantage of faster convergence.

\[ E_{\text{Dihedral}} = \sum_{ijkl} K_{ijkl} \left[ 1 + \cos(n\omega - \omega_o) \right] \] (Eq. 8)

### 1.2.2 Generalized Force Fields

Generalized force fields are a combination of databases, tables, and programs to access topology information and assign atom types, connectivity, and charges based on fragments for automatically constructing force fields for any molecule. They are largely used for initially generating files for further parameter refinement. However, more recent implementations have
had significant initial accuracy for generating molecules with well-studied atom types and connectivity. The most commonly cited generalized force fields are the AMBER (29–31), CHARMM (32,33), OLPS (34–36), and MMFF (37) formats. They each have multiple revised versions continually updated as research advances and are optimized under different models with various datasets and refinement procedures.

Selection of a force field file format is mostly dependent on the software being used for modeling (See Table 1) and does not have an overall effect on the quality of a model, but a generalized force field can influence accuracy due to what type of systems and conditions for which a generalized force field has been optimized. Some examples are Gromos43A1, OPLS-AA, and CHARMM27 being known to be in best accordance with NMR data concerning helical structures while others tend to overpopulate the alpha-region (38). Each vary in the treatment of solvent models, depending on which one they are optimized with. With respect to MD simulation of biomolecules, many force fields have trouble undergoing transitions to non-native conformational states beyond 100 ns simulation time (39). They also differ in ability to form certain secondary structures, such as hairpins. CHARMM27 and OPLS-AA are particularly poor at this (40). However, later implementations of OLPS, OPLS2005 and OPLS3, have made significant strides in overcoming these issues, becoming one of the better generalized force fields to date.

CHARMM has a large contributing user-base and many versions and variations of the force field exist. The main generalized force fields are CHARMM22 and CHARMM27. For protein simulation, they are equivalent, but for nucleic acids, CHARMM27 is optimal.
AMBER force fields are very diverse and optimized for very specific simulations. As of 2018, ff14SB is the current iteration for protein, lipid14 is used for lipids, and ff99 for nucleic acids.

The OPLS-AA, OPLS-2005, and the most recent OPLS3 are generalized force fields optimized to fit the experimental properties of liquids, such as heat of vaporization and density, in addition to fitting gas-phase torsional profiles. OPLS force fields were optimized with TIP3P and TIP4P water models.

The Merck Molecular Force Field (MMFF), derived from computational simulation, is not optimized for one use, such as simulating proteins or small molecules, but tries to perform well for a wide range of calculations on small to medium-sized organic molecules. GROMOS is a parameter set used in the MD software GROMACS and is a united atom force field initially optimized with respect to the condensed phase properties of alkanes (41).

### 1.2.3 Customized Parameterization of Force Fields

Parameters for force fields are constructed in a semi-empirical manner, combining experimental data and small molecule electronic structure calculations. Correcting values for force field parameters with experimental data improves accuracy due to the inclusion of bulk phase properties and phenomena that small-scale \textit{ab initio} modeling methods cannot simulate. Much of the parameterization is done with iterative quantum and classical simulations, using experimental quantities for verification.

Lower complexity parameters with stiff degrees of freedom, like bond stretching and angle bending, are the easiest to fit, often experimental vibrational spectra and structural data are
sufficient, with some electronic structure calculations on smaller molecular fragments for refinement.

Lennard-Jones parameterization is usually done by finding values that reproduce bulk properties, such as density and enthalpies of phase change, in iterative simulations of small molecules and molecular fragments. Initial parameters can be quickly tabulated from van der Waals radii and energetics extracted from critical point data and experimental crystal packing.

Partial charges are then assigned based on *ab initio* calculations of the electron density surrounding small molecular fragments using Mulliken populations (used by CHARMM) or electrostatic potential derived point charges such as Restrained Electrostatic Potential Fit, RESP (used by AMBER and OPLS). Additionally, if a molecule has several relevant conformers (such as rotations around a torsional angle), or external polarization factors exist (like solvent or protein interactions), separate electron densities need to be computed for each conformer. The electron density can then be rendered into discrete atomic partial charges by various methods to reproduce the electrostatic potential around the molecule.

Finally, to account for any remaining energetic contributions needed for agreement with electronic structure calculations not already captured by van der Waals and partial charge interactions, torsional potentials, proper and improper, are used to adjust the energies of rotation around bonds.

To generate these parameters, quantum simulations should be employed with perturbation (MP2), restricted Hartree-Fock (RHF), or hybrid methods between Hartree-Fock and density functional theory (DFT). Basis sets chosen for the geometry optimization should not be coarse-grained and are at least 6-31G, and often require going up to 6-311G**, and may include diffuse
functions depending on the size and charge complexity of the molecule in question. Quantum mechanical calculation is also required for experimental parameterization of small molecules to verify accuracy of a force field. If a generalized force field is shown to be insufficient or inaccurate, a new force field parameterization may have to be generated through extensive QM simulation (Section 1.3.4 Quantum Chemistry). If a protein exhibits significant quantum interactions (i.e. “behaves poorly” in a classical system), a new force field may have to be generated as well.

Table 1. Software that implements particular force fields.

<table>
<thead>
<tr>
<th>Program</th>
<th>AMBER</th>
<th>CHARMM</th>
<th>OPLS</th>
<th>OPLS2005</th>
<th>OPLS3</th>
<th>MMFF</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMBER</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>CHARMM</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>GROMACS</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>NAMD</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Schrödinger</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>YASARA</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

*Force fields are most recent implementations as of this publication.*

### 1.2.4 More Complex Models

Metal coordination, ligand reactions, heme groups, and many other common interactions are very difficult to account for in molecular modeling and are often neglected, resulting in significant decrease in model accuracy. Attempts to overcome or simplify these issues have resulted in more complex force fields with varying degrees of success.
One significant problem with classical force fields is that they are not able to account for induced polarization (rearrangement of charge distribution in a molecule) caused by adjacent interacting molecules. Intra-molecular polarization (polarization of atoms in the same molecule due to conformational rearrangements) should be largely accounted for through proper force field parameterization. Many methods have been conceived for incorporating this effect into classical force fields with two common approaches: perturbation of atomic partial charges or creation of inducible dipoles on each atom. For each conformation of the molecule, the partial charge or induced dipole on each atom must be solved iteratively. The iteration typically reaches a self-consistent state where the electric field due to the charges or dipoles is consistent with the induced response in the molecule (42). Unfortunately, this added iteration during each step of a simulation greatly increases the computational cost when using polarizable force fields, making them expensive to simulate and limiting their use. Despite this difficulty, with future increases in computing power and the need for higher accuracy in simulations, polarizable force fields are beginning to be explored further (43). The CHARMM-based Drude force field (44) and the fragment-based molecular orbital method, Xpol (45), are some examples of current polarizable force field implementations being researched.

Another severe issue with classical force fields is their inability to model chemical reactions due to the dependence on having all bonds explicitly defined. Reactive force fields attempt to replace explicit bonds in favor of bond orders, which can allow for continuous bond formation and breaking. ReaxFF is a bond order-based force field developed at the California Institute of Technology (46). ReaxFF attempts generalization but the more complex parameterization often relegates it to limited context and specific simulations (47). It has been parameterized and tested for many hydrocarbon reactions and high-energy materials, as well as
some more specialized problem such as metal catalyzed nanotube formation and alkoxy silane gelation, and oxygen interactions with non-idealized silica surfaces. ReaxFF relies on high accuracy DFT based on Minnesota Functionals (48).

A number of other force fields have been developed to better capture specific behavior in atomic systems. These methods all aim to improve agreement with quantum mechanical behavior and hence increase quantitative accuracy, but are beyond the scope of this chapter.

1.3 Modeling Methods, Algorithms, and Techniques

Organization of the following techniques are based on a general first pass through molecular modeling. It should be noted that research is rarely linear and often requires multiple iterations to refine a model. As such, this section may refer ahead to other subsections or pass over certain subjects to be elaborated on in more depth later.

1.3.1 Protein and Ligand Preparation

Artifacts, assumptions, and omissions from various structure determination methods are ubiquitous and must be accounted for prior to simulation. Preparation of protein/ligand is required for any program and each software application has its own proprietary method of indexing molecules, chains, residues, and atoms. Some programs use model building scripts or “wizards” to assist in these tasks like the Schrödinger program, Protein Prep Wizard. Some of the more common concerns are side chain modifications, hydrogen treatment, and poor initial atom placements. For X-ray structures, replacing the sulfur atom in methionine with selenium is a regular practice to assist in structure elucidation. Selenomethionines do occur naturally, but converting them back to methionine is needed if done artificially. Treatment of hydrogens is a
particular difficulty in structure determination due to their small size, overwhelming abundance, and rapid motion. Scripts for insertion, deletion, and reordering of hydrogens help account for differential protonation states, insufficient electron densities, and unintentional spatial overlap. Since PDB entries are frequently a single structure averaged from an ensemble of conformational data, this positional averaging regularly leads to overlapping atoms that do not naturally exist in such a high energy state (49). Careful energy minimization should be used to slowly push a system to a less volatile state.

1.3.1.1 Determination of Protonation States. As mentioned previously, hydrogens are a pervasive nuisance in structural determination causing very few crystal structures to possess atomic resolutions accurate enough to assume the protonation state of a protein. Conditions required for crystallization also often require non-physiological pHs leading to unphysical protonation of side chains. Empirically-based programs, like PROPKA (50), can estimate the initial protonation state of side chains. Other physics-based methods exist for finer pKa estimations using classical Poisson-Boltzmann or quantum mechanical free energy calculations. Since proteins are dynamic entities, protonation states can change throughout simulation due to conformational shifts inducing local environmental changes on side chains. Key residues, histidine for instance, that are susceptible to alternate protonation states (i.e., tautomeric states) should be identified, and iterative and/or separate simulations are recommended.

1.3.2 Energy Minimization

Energy minimization is often crucial on prepared protein structures and ligands and should be done individually so as to not bias conformations towards initial structures. This
allows for any algorithm to rely on its intended function and helps to limit false model convergence. Energy minimization is necessary to alleviate overlapping atoms, close contacts, or unrealistic constraints imposed by structural resolving methods. Use of constrained energy minimization, with successive relaxation of these constraints, is useful in this context in order to prevent unrealistic divergence from the crystal structure during the relief of close contacts.

1.3.2.1 Energy, Entropy, Free Energy, and Enthalpy. The concept of energy itself is ambiguous and does not physically exist. Energy is simply the sum of all possible forces, kinetic (T) and potential (U), in a system and is useful for understanding and modeling more complicated systems. This is also known as the Hamiltonian definition where the energy term is H. For modeling dynamics, the principle of least action is implemented through the Lagrangian, L.

\[ H = T + U \quad \text{Hamiltonian} \quad (\text{Eq. 9}) \]

\[ L = T - U \quad \text{Lagrangian} \quad (\text{Eq. 10}) \]

Which forces are deemed relevant for inclusion into any heuristic, ad-hoc definition of energy is largely arbitrary, based only on what may be important to a researcher at the time. This almost always leads to confusion when energetics are employed to explain certain observations or predict other quantities. For biochemistry, focusing on thermodynamic quantifications of energy tends to be most useful with the Grand Canonical Ensemble to represent the possible states of a mechanical system of particles.
\[ U = TS - PV + \mu N \]  

(Eq. 11)

Where \( U \) is the total internal energy of a system with the components \( TS \), temperature (\( T \)) and entropy (\( S \)), representing the amount of heat in a system, \( PV \), pressure (\( P \)) and volume (\( V \)), denoting the work done by or on the system, and \( \mu N \), chemical potential of a single molecule (\( \mu \)) and number of molecules (\( N \)), inferring the total potential for work to be done by or be extracted from the system. Each of the multiplied constituents in the equation demarcate a specific subdivision of potential and kinetic forces in a system.

Expansive and contracting motions are defined by work, \( PV \). This term does not hold much practical use for biochemistry and is rarely implemented since most experiments are carried out under constant pressure.

The more valuable terms are heat and total chemical potential in a system. Heat is a measure of unavailable energy in a system locked up in conformational or structural modes. This term is a composite of the average relative kinetic energy of a system, otherwise known as temperature, and the possible number of states accessible to the system, or entropy, at a given temperature. Entropy is nearly impossible to measure directly but is an extremely valuable term describing how energy is stored in a system.

The last term, \( \mu N \), is arguably the most important. It expresses the amount of work a system could perform or free energy. Most calculations attempt to derive or predict values associated with free energy since it can elucidate phenomena that can or cannot occur in a system given a particular state. Unfortunately, free energy is difficult to extricate from other energy terms. Empirically, enthalpy, the total chemical energy and heat contained in a system, is an easier value to measure. For example, the enthalpy of formation is the total energy required to
create all bonds and configurations in a molecule. By measuring the heat (TS) lost and work (PV) done when all bonds are broken, the free energy of a bond can be deduced from the remaining energy.

Theoretical calculations of free energy are just as challenging. Many procedures rely on perturbatory methods on equilibrated systems. For understanding potential effects on protein binding with differing ligand, alchemical transmutation is advantageous where the atoms in a ligand are transformed into different elements or molecular moieties. Any change in the equilibrated system’s energy must be caused by the new functional group. For conformational or positional effects, umbrella sampling methods allow for free energy determination by fixing a predefined reaction coordinate in various states and measuring the associated energy changes. Thermodynamic integration and Metropolis Monte Carlo methods, which are not based on perturbation, are also available for more specialized systems utilizing clever exploitation of various thermodynamic quantities. Each system requires a novel approach for free energy determination, so no algorithm can be made to automatically setup a system for ascertaining this measure. Forming an appropriate hypothesis and testing environment for extracting explicit energetic terms demands rigorous attention and consideration of all assumptions and constraints. Again, discussing energy alone is meaningless, so any use of energy terminology must include specific definitions of the quantities examined.

1.3.2.2 Energy Minimization Algorithms. Many computational methods exist, but the Newton-Raphson method is one of the most accurate and commonly used algorithms. Based on a Taylor series expansion of the potential energy surface at the current geometry, it is the most
computationally expensive per step of all the methods utilized to perform energy minimization. However, it usually requires the fewest steps to reach the minimum (51).

For rapid estimation, the steepest descent algorithm relies on approximations instead of requiring explicit calculation of numerous second derivatives. The second derivative is assumed to be a constant making it much faster per step than the Newton-Raphson method. This computational gain is unfortunately offset by its inefficiency, and more steps are generally required to find the minimum if it can be found at all (52).

The conjugate gradient method is a decent compromise between speed and accuracy. Similar to steepest descent, the search first takes place in the direction of the largest gradient. In contrast to steepest descent, the conjugate gradient method remembers the direction of the previous search to help avoid some of the persistent oscillating behavior that often plagues the steepest descent as it moves toward the minimum. This gives a greater likelihood to find a minimum with far fewer iterations. For larger molecules, the expense of the Newton-Raphson method becomes even more pronounced, leading to a much higher computational cost than other methods. Despite the smallest amount of CPU time per step of the three methods, steepest descent requires many more steps to find a minimum due its overall inefficiency. As such, the conjugate gradient method is the most commonly used method for energy minimization of large molecules (53). The Truncated Newtonian Conjugate Gradient (TNCG) adaptation of the original conjugate gradient method is a far faster variant implemented in most energy minimization codes (54).

These energy minimization methods do not incorporate any dynamics and seek only to lower local potential energy. Biomolecules have evolved to rely on thermal vibrations to maintain an optimal structural equilibrium. Multiple minimization runs will drop a structure into
a local energy well further than thermal kinetics would allow. This over-minimization is analogous to freezing and is not physiologically relevant to biomolecules (55). For exploring global energetic minimums, other minimization protocols, such as simulated annealing (Section 1.3.3.2 Simulated Annealing), can be used to overcome the locality problem in a molecule’s potential energy surface.

1.3.2.3 MM/GBSA. The procedure haphazardly named Molecular Mechanics energies combined with Poisson–Boltzmann (MM/PBSA) or Generalized Born (MM/GBSA) and Surface Area continuum solvation is essentially a minimization procedure designed to estimate free energies of binding. By analyzing multiple conformations, normally through MD or MC simulation, free energy estimations are obtained from energy minimizations on the receptor, ligand, and complex individually and combined with the following equation,

\[ \Delta G_{\text{bind}} = \Delta G_{\text{complex}} - \Delta G_{\text{receptor}} - \Delta G_{\text{ligand}} \]  

(Eq. 12)

While MD is not necessarily required for a large conformational search set (56–59), it is highly recommended in order to capture dynamic processes and for statistical validation of the approach (60). MM/GBSA also requires knowledge of the solvent and electrostatic environment around the target system.

To account for solvation effects when the atomic details of the solvent are not relevant, a continuum approximation can be used to reproduce thermodynamic properties of the system. By rolling a coarse-grained solvent molecule over the target system, the solvent accessible surface
area is traced out and calculated allowing estimation of the free energy required to move a molecule from an aqueous solvent to a non-polar one.

Electrostatics can be handled by solving the Poisson-Boltzmann equation. Though computationally demanding to solve, it gives good estimates of the electrostatic component of the solvation free energy. The overall quality of the approximation relies on the solvent behavior being reasonably close to a dielectric continuum. For bulk water, this is a very good estimate; for interface waters or where molecular granularity is important, such as water bridges, this approximation breaks down quickly. An alternative algorithm for estimating electrostatics is the Generalized Born (GB) method, a fast, highly accurate, approximate solution to the Poisson-Boltzmann equation (61).

MM/PBSA and MM/GBSA are useful for improving the results of virtual screening and docking and the approaches can reproduce and rationalize a number of experimental findings. However, owing to the numerous and nested approximations and assumptions, performance varies strongly with the tested system and can be difficult to correlate free energies with experimental values. Additionally, many attempts to improve accuracy with a more refined method such as quantum calculations, polarizable force fields, or improved solvation models have actually deteriorated results (62–64).

GB was originally developed by Clark Still (65), and then incorporated into the MM/GBSA method, being later adopted by many other software platforms including GROMACS, NAMD, and Schrödinger (66–68).

### 1.3.2.4 Low Frequency Normal Mode Analysis.

Since protein function is almost entirely predicated on dynamic motions, analytical methods pursuant to the understanding and
categorization of protein flexibility can yield significant insight into the biological and evolutionary purpose of a protein. Normal mode analyses (NMA) look to classify and rank collective protein motions into independent (i.e. normal) harmonic oscillatory behavior (i.e. modes). A normal mode is a concerted motion of many atoms where the center of mass does not move and all atoms pass through their associated equilibrium positions at the same time. Normal modes are also directly related to vibrational spectroscopy (e.g. IR). NMA methods apply a great deal of assumptions that sharply hinder the method’s general use and requires explicit equilibration of the initial system (see section 3.3.3 Equilibration). NMA assumes that all pairwise forces behave like springs for short displacements, all motion vectors are linear, and all fluctuations are in a ground state harmonic energy well with force constraints able to be approximated as the second derivative of the energy multiplied by the displacement in a specific direction. The assumptions allow the system to be expressed as a Hessian (2nd derivative) matrix where the normal modes are eigenvectors of the matrix. Closed forms for the second derivatives are rarely available, so they are calculated numerically.

The general procedure for NMA is initial structure energy minimization and subsequent Hessian generation followed by diagonalization of the Hessian to solve for the eigenvectors and eigenvalues. Eigenvalues correspond to amplitude of a motion. Low eigenvalues are associated with low force constants or low frequencies. Eigenvectors denote the directions of movement and are valuable for visualizing the collective movements. These quantities can also be extracted from an MD simulation with concomitant expense in computational time (69). It is not known which modes are relevant for protein function but extensive surveys of structural databases have shown that structural transitions of many proteins can be explained by just a few of the lowest-frequency normal modes (70). Therefore, most NMA is only concerned with low frequency
vibrations (5-10 lowest eigenvalues), commonly abbreviated as low-mode (71). Low-modes are similar to atomic normal modes constructed from crystallographic models and can be beneficial in extracting information on protein function from poor resolution X-ray or NMR structures and homology models since they require little resolution to reproduce functional rearrangements, even as low as 30Å resolution(72,73).

1.3.3 Molecular Dynamics and Monte Carlo

Molecular Dynamics (MD) is a classical simulation of the motion of molecules by solving Newton’s equations of motion for a virtual system with respect to time, temperature, and pressure. Monte Carlo (MC) methods are a broad class of computational algorithms that obtain numerical results by repeated random sampling. The fundamental concept in MC is to employ randomness for solving problems that are likely deterministic in principle. MC generates states according to Boltzmann probabilities where MD attempts to reproduce the dynamics of a system. A Markov chain procedure is used to determine a new system state from a previous one. This new state is accepted at random, using certain criteria such as a Bennett acceptance ratio, to keep the system stochastic.

1.3.3.1 MC vs MD. MC is usually superior for conformational sampling as large energy barriers, often several kₜT, exist for torsional rotations. This tends to cause poor state sampling in MD simulations due to molecules being trapped in a few low energy conformations. In contrast, the random moves in a MC simulation can easily produce barrier crossings if the energy window for accepted moves is sufficient. Unfortunately, this strength of MC can also be a difficulty for the procedure and cause for error (74). Choosing an energy window depends on the size of a
system and the conformational behavior being studied. Leaving the energy window near $k_B T$ will likely not produce any significant state changes but is helpful in elucidating local degenerate structures. Larger windows, above 100 kcal/mol, will generate a very diverse set of states, but for proteins this is often useless unfolding behavior. Biomolecule energy windows around 5 or 6 kcal/mol are ideal since that is near the thermal energy contained in bulk water at body temperature (75).

MD becomes favorable for liquid simulations where molecular collisions exchange energy between molecules, enabling barrier crossings, and thus improving the ability of MD to sample system conformations. MD may also be more costly in terms of computation, but in MC simulations, there exists a high chance of selecting random moves where atoms or molecules overlap, especially for rotations of nematic molecules with long tails such as phospholipids or liquid crystals. This results in a large number of rejected moves and subsequent decrease in sampling efficiency. For larger systems, this issue often negates the lower computing time of MC, frequently making MD faster overall. Despite this, the ability of MC to make unphysical moves can compensate for this slow down in some cases, breaking through energy barriers that may take MD an inordinate amount of time to sample. Moreover, MD handles collective motions in general better than MC, though recent methods such as Hybrid Monte Carlo and Configurational Bias Monte Carlo have been developed to improve the performance of MC simulations in these areas (76). MacroModel (77), MCPRO (78), and the open-source FEASST (79) are some software programs available for MC molecular modeling simulation. While many systems can use either method to achieve the same result, some applications can only use one method. Determining transport properties, such as viscosity coefficients, is only possible with MD, since MC lacks any objective definition of time. On the other hand, MC can be used for
simulations with varying numbers of particles, a procedure known as Grand Canonical Monte Carlo, by defining moves for the destruction and creation of particles. For the most part, the avoidance of dynamics relegates MC to the study of static quantities only (80).

Table 2. Table of available MD Software

<table>
<thead>
<tr>
<th>Program</th>
<th>Replica Exchange</th>
<th>Simulated Annealing</th>
<th>Free Energy Perturbation</th>
<th>Implicit Solvent</th>
<th>GPU Accelerated</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMBER</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>CHARMM</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Desmond</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>GROMACS</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>NAMD</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

*Indicates if a particular program has a preexisting algorithm or interface for a certain procedure requiring minimal effort for system setup.*

1.3.3.2 Simulated Annealing. One of the advantages of MC compared to MD is the method’s ability to surmount energy barriers (81,82). Simulated annealing (SA) can confer some of this capacity to stochastic dynamics for overcoming large energy wells. In terms of molecular systems, SA is the process in which a molecule is heated progressively to 600+ K with a very small time step (less than 1 femtosecond) for a very short length of time (about 10 picoseconds). This can “kick” a molecule out of possibly artificial energy wells and unreasonable conformations. The temperature is then gradually lowered back down to more relevant temperatures. The rapid heating opens up more possible states for a systems, but the entropy incurred by this process often yields even more unreasonable conformations. A well-planned annealing schedule is important for slowly lowering the system into the global minimum. Many SA runs tend to be required for this followed by energy minimization once more after the
annealing is finished to allow the structure to properly settle into the local energy well. This stepwise process gives a relatively quick assay of stereoisomer conformations, optimal dihedrals, bond lengths, and molecular geometries.

1.3.3.3 Equilibration. The discipline of statistical mechanics is intrinsic to all computational modeling with the axiomatic concept of equilibrium at its core. A system at internal thermodynamic equilibrium observes no macroscopic change even when microscopic thermodynamic variables fluctuate. Therefore, any collective shifts in the average thermodynamic quantities can only be due to external influences imposed on the system. Without this global assumption of equilibrium, determination of cause and effect for an isolated variable cannot be verified. For this reason, great care should be taken to ensure a system has convergence around appropriate constants (Section 1.3.3.5 Trajectory Analysis for Equilibration). A system should eventually evolve into a state that is independent of any history and simulation time. This is referred to as equilibration. If it does not, the system cannot be adequately described by computational modeling and demands further examination of initial assumptions and conditions used to build and/or simulate the dynamics of a system (83).

It is important to note that thermodynamic equilibrium is not the same as dynamic equilibrium. Thermodynamic equilibrium is achieved when thermodynamic quantification of a system is stable and is no longer influenced by its initial configuration. The thermodynamic quantities, such as pressure, temperature, and volume, are not changing above thermal fluctuation and the potential energy of the system is minimized for a local energy well. This does not mean that total energy is constant. Often, the total energy of the system appears constant for short timescales since all systems equilibrate thermodynamically into some local energy well and
will oscillate within it until a thermal fluctuation kicks the system out into another energy well. Dynamic equilibrium is achieved when the system reaches its global structural energetic minimum.

1.3.3.4 Common Setup Parameters for MD Simulation. MD simulations use a periodic boundary box to spatially constrain the simulated system. Many enclosure geometries exist, though cubic boxes are often a better choice for simplicity, but an orthorhombic box can take more advantage of the shape of a protein or ligand to minimize box volume. Decreasing box volume reduces the number of solvent molecules needed to encase the solute and thereby lower the number of calculations entailed for the simulation. These two geometries are also almost always compatible with any MD software. Proper periodic boundary conditions can be established by allowing at least 5-10 Å of buffer from the solute to the box boundary. This is where MD commonly has issues with unrealistically imposed high solute concentrations as bordering boxes can affect each other through long range electrostatics. Ensuring interface waters, solvent affected by the solute, do not touch each other by increasing the distance of the solute from the box boundary will greatly reduce this self-interaction. Bulk solvent occurs when the dissolved solute no longer influences the solvent. Distinct metrics should be asserted to test for bulk properties in the solvent to prove adequate buffering. For a constant temperature and pressure, density is entirely dictated by the intermolecular forces described in the force field file making it a simple and sufficient indicator of force field accuracy. A protein touching its periodic image is completely unphysical, akin to shaking your own hand through a mirror. The choice of a solvent model is another influential decision on an MD simulation. Water is the most common solvent and the most difficult to simulate classically. The GROMOS force field is optimized with
the Simple Point Charge (SPC) and SPC/E model. OPLS and CHARMM are optimized with TIP3P and TIP4P. Multi-site models may be necessary if protein or ligand displays a dependence on water or has coordinated waters. Many other solvent models exist, such as chloroform or octanol, but may not be available for some MD model builders.

Table 3. Overview of molecular timescales for general atomic and protein movements.

<table>
<thead>
<tr>
<th>Motion of Interest</th>
<th>Approx. Number of Atoms</th>
<th>Distance Amplitudes</th>
<th>Average Timescales</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bond Stretching and Angle Bending</td>
<td>2 – 3</td>
<td>0.001 – 0.1 Å</td>
<td>1 – 5 femtoseconds</td>
</tr>
<tr>
<td>Side chain and dihedral motions</td>
<td>6 – 20</td>
<td>0.1 – 2 Å</td>
<td>1 – 10 picoseconds</td>
</tr>
<tr>
<td>Loop and Collective Protein Motion</td>
<td>50 – 1000</td>
<td>1 – 100 Å</td>
<td>10 – 100 nanoseconds</td>
</tr>
<tr>
<td>Protein Domain Motion</td>
<td>100 – 10000</td>
<td>10 – 100 Å</td>
<td>0.1 – 5 microseconds</td>
</tr>
<tr>
<td>Protein Folding</td>
<td>1000 – 100000</td>
<td>10 – 500 Å</td>
<td>1 – 1000+ microseconds</td>
</tr>
</tbody>
</table>

Estimating an acceptable runtime for achieving equilibrium is not trivial. The time span simulated should match the kinetics of the natural process, implying that an initial understanding of the system is necessary. For many small (~<50 residues) stable proteins with accurate crystallographic data (~<2Å resolution with no missing or unresolved residues), the minimum time required is at least 10-20 ns. This is, again, a minimum and larger systems (10,000+ atoms) with greater degrees of freedom require longer simulations (Often 100-1,000 ns). Table 3 has a synopsis of some basic kinetics for protein systems (84). For simulations longer than 100ns, special care must be taken to refine force fields to adequately reproduce known structures since most generalized force fields do not possess the required precision and systems start to deviate significantly past this scale (39).
Classical MD simulations follow an approximate \( O(N^2) \) rule. The “big O” notation is a function of time complexity for the upper bound of an algorithm where, in this case, \( N \) is defined as the number of atoms. This means that if the number of atoms in a system doubles, then the computation time quadruples. On average, with 4 CPU cores and a 30,000 atom system, computation time is about 0.5 nanoseconds simulated per day. This makes GPU acceleration almost mandatory for MD, reaching hundreds of nanosecond per day for an equivalently sized system. As such, nearly all current MD programs are optimized for running on GPUs.

The thermodynamic ensemble applied to simulation is commensurate to the scale of the phenomena being studied. The adiabatic Microcanonical (NVE - constant number, volume, and energy) ensemble is implemented when the total energy of a system must be conserved and is useful for simulating chemical processes, even though simulation time and temperature are not well defined under this constraint scheme. QM simulations are usually run in a Microcanonical ensemble. The isothermic, isochoric Canonical (NVT - constant number, volume, and temperature) and the isothermal, isobaric Grand Canonical ensemble (NPT - Constant number, pressure, and temperature) are closer to experimental configurations, though it is not a minor task obtaining a Canonical distribution of conformations and velocities. Dependency on system size, thermostat choice, thermostat parameters, time step, and integrator is the subject of many articles in the field. For MD simulation, the differences between NPT and NVT are usually quite small (85–87). Therefore, NPT is a better choice as this method corresponds most closely to laboratory conditions with a flask open to ambient temperature and pressure. However, if the simulated system is very large and NPT is presenting difficulties with equilibration, especially if the system exhibits significant dispersion effects, leading to unstable density reduction, then NVT may be a better ensemble choice. In the simulation of biological membranes, isotropic pressure control is
not appropriate. For lipid bilayers, pressure control occurs under constant membrane area (NPAT) or constant surface tension, gamma (NPγT).

For defining a thermostat and barostat, the Nose-Hoover thermostat for NPT simulation is recommended. Temperature choice should be based on the experimental methods being used to correlate simulations. Biological temperature (310 K) or room temperature (298 K) are the most common temperatures for protein systems. The Parrinello-Rahman barostat is advised when Nose-Hoover thermostat is used, which works by changing the box shape (i.e. volume) to affect the simulation pressure. Martyna-Tuckerman-Klein (MTK or MTTK) is an extension of the Parrinello-Rahman barostat and is often preferred for protein simulations.

With regard to simulation timesteps and integrators, the Reversible rEference System Propagator Algorithm (RESPA) is favored since it is one of the more stable and empirically validated integration schema. The Verlet algorithm is another common and stable integrator, recommended for NVE simulations. A timestep of 2 femtoseconds (fs) is advised for rigid bonds, 1 fs for flexible. Raising the timestep can reduce computational loads, but is broadly discouraged as it usually causes severe artifacts at best and simulation explosion at worst.

Long-range electrostatics are the most computationally expensive calculations in an MD simulation with Particle Mesh Ewald (PME) (88,89) being one of the more common and effective correction methods. The reaction field (RF) (90) method is another solution that yields the same accuracy as PME, but with slightly less computational costs (91,92). Imposing cutoffs can greatly reduce computation time but at a cost of potentially inducing significant error over the course of a simulation. A 20Å cutoff is advocated but is usually not feasible due to size and system constraints. For proteins, a minimum of 13Å is recommended, though cutoffs down to 9Å can still be adequate. Small molecules can get away with even smaller cutoffs. Potential and
force tapering are also applicable for amending short- and long-range electrostatics but are not as effective in error reduction as having longer cutoffs (93).

Positional constraints may also need to be imposed on a system to help approximate larger structural restrictions or simulation conditions. Restriction of motion is correlated to the thermal energy of the solvent. The thermal energy contained in bulk water at 298K is about 5 kcal/mol. Rigid restriction is any energy value much greater than this, thermal restriction is near this value, and loose restrictions are for energies much less than solvent’s thermal energy. Any bias imposed on the system drastically affects the physical relevance of the simulation. Therefore, constraints should be carefully reviewed and correlated to known systems.

1.3.3.5 Trajectory Analysis for Equilibration. Equilibration is the primary goal in MD simulations. Without it, no other quantity can be confidently determined. Many different general principle components can be analyzed to establish that system equilibrium was achieved. The more popular ones are Root Mean Square Displacement (RMSD), Root Mean Square Fluctuations (RMSF), and Radius of gyration (R_g). RMSD is the most prevalent analysis for measuring equilibration. It quantifies the difference in inter-atomic position over time and can help to catch possible conformational shifts in proteins. Root Mean Square Fluctuation (RMSF) is essentially the RMSD per residue averaged over the length of the simulation. X-ray crystallography uses a metric called B-factors related to RMSD and RMSF.

\[
B = 8\pi^2 \langle \text{RMSD}^2 \rangle \Rightarrow \text{RMSD} = \sqrt{\frac{B}{8\pi^2}} \approx \sqrt{\frac{B}{79}} \quad \text{(Eq. 13)}
\]
This equation roughly shows that an atom with a B-factor around 80 will be localized in a sphere with a radius of 1Å about half the time. Analysis of individual residue B-factors identifies the most flexible regions of the protein. For a successful simulation, secondary structures in proteins must be maintained and core residues should not move beyond thermal fluctuation (~ +/- 1Å). Radius of gyration \( (R_g) \) calculates the centroid of protein and average intra-atomic distance to that centroid. This is useful in measuring density shifts but is generally only meaningful for globular proteins since many geometries can have identical centroid positioning. Alpha carbon distance maps are another good tool for large oscillatory behavior. A principle component analysis on multiple alpha carbon distance maps can be correlated to RSMF. Normal mode calculations (Section 1.3.2.4 Low Frequency Normal Mode Analysis) may be useful for verifying oscillatory behavior but, since normal mode calculations require all interactions to be constant, this method has limited use.

These metrics supplicate a comprehensive review to identify if and when equilibrium was achieved. Analyzed properties such as RMSD should converge or asymptote/plateau to a reasonable average value (A good RMSD cutoff is near the PDB resolution). Multiple plateaus may exist for longer run times. Final analyses should only be performed on the equilibrated frames and no pre-equilibration frames should be included into production. Production runs should be at least 5 – 10 times longer than the equilibration times to properly investigate the local energetic minima. Finally, conclusions should be consistent with the time scales of the simulation (See Table 3).
1.3.4 Quantum Chemistry

Energy minimization, MD, and MC are all deterministic classical methods that do not treat electrons and nuclei as separate entities nor account for uncertainty in their motions. As such, these methods cannot describe high resolution details of atomic interactions. This is most evident in the inability to characterize chemical reactions, but also leads to artifacts and poor replication of physical phenomena. Since all physical interactions are governed by quantum mechanics (QM), discretization of fundamental quantities described by probabilities as opposed to continuous functions with definite precision, classical interpretations are often inadequate. QM methods are needed to elucidate and quantify many chemical phenomena, such as reaction paths, spectroscopy, and the prediction of many atomic and molecular properties. QM computation is vital throughout many disciplines, not just in the more directly related sciences of physics, chemistry, and engineering, but also in biology with respect to drug discovery and molecular biochemistry.

As described previously (Section 1.2 Selection of Molecular Mechanics Force Fields), quantum mechanical (QM) computation is needed for the creation of force fields in classical simulation, but is also commonly required to further refine the initial geometry of a system. Some classical iteration is still needed before running QM simulations as they are also extremely dependent on the initial configuration of the system. All previous preparation steps (protonation state analysis, MD, minimization, etc.) are advised and may have to be done iteratively until sufficient convergence. Even though QM simulation is far more accurate than classical methods, it possesses extreme computational cost. The total computational time is dependent on the level of theory and accuracy desired. QM calculations (for DFT) follow an approximate $O(N^3)$ rule. This means that if a simulation of 100 atoms takes a week to run on one core, then a simulation
of 300 atoms will take approximately 27 weeks to complete \((3 \times \text{number of atoms} = 27 \times \text{simulation time})\). Proteins are very large molecules (thousands to hundreds of thousands of atoms) and are therefore often impractical to run simulations with greater than 300 atoms without incredible computational resources. If a relatively small 3,000 atom protein were run using the same configuration described previously, it would take about 520 years to finish the QM simulation on the same number of cores. This example is greatly simplified as for \emph{ab initio} calculations, \(N\) actually stands for the number of basis functions, not the number of atoms. Each atom can have as many or more basis functions then the number of electrons, so larger elements contribute even more computational cost to the final simulation.

It is worthwhile to note that QM geometry optimization is fundamentally different from MD simulation. MD simulates the motion of molecules with respect to time subject to initial velocities, temperature, ensemble pressure with solvent, and other approximated meso to macroscopic effects. This permits the MD time evolved trajectories of the system to have some physical interpretation. Geometry optimization, by contrast, works much the same way as classical energy minimization by reducing the forces acting on each atom in a system. The pathway with which it achieves this lower energy state lacks any physical meaning. Optimization algorithms are path independent and can yield the same energy minimized structures.

\textbf{1.3.4.1 Quantum Calculation Methods.} Solutions to the Schrödinger equation that describes the wave function of a quantum system are rarely amenable to analytical means due to \(N\)-body coupling interactions. As such, numerous numerical and approximative methods have been developed to describe the evolution of atomic systems on quantum scale.
One of the first and more successful approximative schemes created soon after the discovery of the Schrödinger equation is the Hartree-Fock (HF) method that uses a self-consistent field (SCF) to estimate charge distributions. HF imposes many strict assumptions that often preclude its unitary use, the worst being a complete disregard for Coulomb correlation (the instantaneous correlation among the electrons arising from their mutual Coulombic repulsion) leading to the inability of HF to describe London dispersion forces. These forces are critical to most macroscopic phenomena for both atoms and molecules. Consequently, other methods are employed to bolster the accuracy in describing many electron systems after initial non-relativistic HF calculation, known collectively as post-HF methods (94). These methods attempt to include some level of estimating electron correlation to the multi-electron wave function. Some methods attempt to expand the wave functions into a linear combination of Slater determinants. The coupled cluster (CC) method, especially CCSD(T), is considered a gold-standard for quantum chemistry. However, this accuracy comes at a very high computational cost, limiting CC methods to relatively smaller systems compared to other methods (95–97). Configuration Interaction (CI) is a linear variational method that confers a unique ability to generalize its formalism to many disparate system configurations from excited states, open-shell systems, and to systems far from their equilibrium geometries (98,99). By contrast, traditional single-reference perturbation theories and CC approaches generally assume that the reference configuration is dominant, and tend to fail when that is not the case. The major downside to CI is that truncation of CI space is necessary for it to be computationally tractable and causes the method to have inconsistent energy behavior with increasing distances and no longer scales correctly with an increasing number of electrons. Other post-HF methods are perturbative, such as Møller-Plesset (MP) perturbation (100), and are relatively cheaper to compute than others if kept at a lower level of
theory while retaining reasonable accuracy. The simpler and less computationally expensive MP2 method gives superior results compared to CI when the size of the system increases. Unlike CC and CI methods, MP perturbation suffers from convergence and accuracy issues that higher levels of theory don’t seem to be able to solve (101).

An alternative to HF that can produce properties of a molecular system based solely on electron density is Density Functional Theory (DFT). DFT was not originally considered a viable alternative to HF methods until recent advancements sufficiently improved the accuracy of the exchange and correlation interactions. It still has issues describing some important molecular properties, such as van der Waals forces and excited states, but the simplicity and computational efficiency of the method provides a good balance between accuracy and computational cost. The method scales as $O(N^3)$ or better, enabling calculations with thousands of electrons. DFT uses functionals to calculate electron densities and probability distributions. Most of the more popular functionals include some HF input to determine exchange interaction energies. The choice of a functional to incorporate into a QM calculation is determined by which functional performs best for a given system. As an example, B3LYP is the standard but gives significant and consistent error and MO6-2X (102) is more accurate for molecules containing atoms smaller than argon but yields substantial error for transition metals, whereas MO6 has been validated as a preferred choice for organometallic compounds (103,104). Many review articles exist discussing benchmarks and optimization of system configurations with the appropriate functional. With over half a century of research behind the theory, DFT is now a very well-established technique with its capabilities and limitations reasonably understood and an abundance of efficient implementations available.
1.3.4.2 Basis Sets. QM simulations are highly dependent on the level of theory used. The theory is further decomposed into the required functional and basis set. In quantum chemistry, a basis set usually refers to the set of single particle functions used to build up molecular orbitals. The term gets confused sometimes by theorists who might also refer to sets of Slater determinants as N-electron basis sets, which is something else entirely (105). Atomic orbitals represented by atom-centered Gaussian functions (GTO) are the prevailing basis set. Some older programs use Slater functions (STO) which have the correct short-range and long-range behavior but GTOs are much easier to compute through the Gaussian product theorem. The Pople basis sets are the most popular with split-valence double-zeta basis set is called 6-31G as its flag-bearer (106). STO-3G is a deprecated minimal basis set in which each atomic orbital is represented by only 3 Gaussians that attempt to mimic the behavior of a Slater function.

For further set refinement, polarization functions can be added demarcated by an asterisk for a function that covers only heavy (non-hydrogen) atoms or double asterisk for hydrogen inclusion. For anions, Rydberg states, and very electronegative atoms with a lot of electron density, like fluorine, diffuse functions are needed to accurately reproduce polarizabilities or binding energies of dispersion forces (107). These diffuse functions are designated by a plus sign next to the basis set. The Dunning correlation consistent basis sets are optimized using CI wavefunctions designed to converge smoothly as more Gaussians are added. These basis sets are optimal if a calculation does not define core electrons as frozen in post-HF computations like MP2, CISD, CCSD(T) but are overkill for simpler HF methods. In general, the more Gaussian functions the more accurate the calculation.
The process of selecting the proper level of theory, accuracy, and convergence threshold is not well defined. Intensive and rigorous empirical evidence is necessary to validate any QM simulations.

Table 4. List of quantum chemistry software

<table>
<thead>
<tr>
<th>Program</th>
<th>License</th>
<th>Semi-Empirical</th>
<th>HF MP</th>
<th>HF CC</th>
<th>HF CI</th>
<th>Post-HF DFT</th>
<th>GPU Accelerated</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMBER</td>
<td>Free</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>GAMESS*</td>
<td>Academic</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Gaussian</td>
<td>Commercial</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Jaguar</td>
<td>Commercial</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>NWChem</td>
<td>Free</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Q-Chem</td>
<td>Commercial</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

*GAMESS software company split into three independent companies with different softwares, the commercial US and UK versions and the open source Firefly version.

1.3.5 Quantum Mechanical Molecular Modeling (QM/MM)

As stated previously, most biochemical systems are far too large to be simulated at any level of *ab initio* quantum theory. At the same time, classical molecular mechanics force fields are not sufficient for modeling processes, where chemical bonds are formed or broken. To overcome these limitations at either extreme, methods have been developed to treat a small subset of a system at a high resolution level of quantum chemistry (QM), while retaining the computationally cheaper molecular mechanics force field (MM) for the larger partition. This approach allows for the study of chemical reactivity in large systems, such as enzymes, or covalent drug or residue interactions.
The main advantage of hybrid QM/MM methods is their computational efficiency. The calculation time of doing classical MM simulations scales as $O(N^2)$, with $N$ as the number of atoms in the system. The majority of the calculation cost comes from computing the electrostatic interactions term in which every particle interacts with every other particle. The use of a cutoff radius, periodic pair-list updates, and recent variations of the PME method have been able to reduce the time complexity between $O(N)$ and $O(N^2)$ (108). That is to say a system with double the number of atoms would take between double and quadruple the computing power to simulate. The other route of QM scales at best as $O(N^3)$ for DFT or even as bad as $O(N^7)$ for very high level coupled cluster calculations, where $N$ is the number of basis sets. Restricted Hartree–Fock calculations can scale as $\sim O(N^{2.7})$, but the accuracy sacrifice is not usually worth it at these scales. Generally, QM regions are small and typically isolated, so the increased computational cost of QM is mitigated by only making the part of the system that is of interest to be treated quantum-mechanically, such as the active site of an enzyme, and the remaining system is treated classically. Further speed increases come from using semi-empirical methods to approximate certain computationally costly parts of the QM calculation as experimentally derived constants.

While QM/MM seems a superior method, implementing it correctly as to not sacrifice the accuracy of QM or the speed increase of MM is a difficult and sometimes insurmountable problem. (109) The most prominent issue is dealing with electrostatics and how to transfer charges between the QM and MM regions. In QM, atoms are not a well-defined quantity in a molecule as there is no quantum mechanical operator that corresponds to an atom. Because electrons are defined by Gaussian distributions, this delocalized behavior leads to ambiguity in the definition of an atomic charge which is critical for MM force fields. Some procedures use population analysis schemes to essentially sum over all electron density using the basis functions
of a given atom to yield a partial charge, known as Mulliken charges. This scheme, however, depends on how the atoms are centered and breaks down as the basis functions become more delocalized. Another scheme is to define atoms as “critical points” of the charge density. This procedure is more stable than Mulliken/Lowdin schemes with respect to basis set expansion. This assists in convergence when more basis sets are added to the system, but it is not clear if this means it is a correct charge design. This particular scheme is implemented in the QM program Gaussian (110). The more common charge expression scheme is ElectroStatic Potential (ESP) fitting where charges are determined by fitting charges which best reproduce the electrostatic potential generated by the molecule. Despite this method’s good fit with MM potentials, multiple solutions for the fit exist and many are clearly unreasonable for the system, especially for the interior of a molecule. Restricted ESP-Fitting (RESP) attempts to avoid these unphysical solutions of ESP charges. The most difficult charge embedding method is covalent embedding. This arises when the QM region of a system is linked to the MM region by covalent bonds. The covalent embedding procedure cuts the bond at the QM/MM threshold and adds two “link” atoms to the system. A hydrogen is added to the newly generated QM “molecule” and a single simplified atom type is bonded to the MM region leading to an overlap of both regions. This method is almost always required in biological contexts since proteins are just very large molecules. Potential problems emerge with this link atom concept as extra degrees of freedom need to be removed, i.e. the link atom somehow needs to be connected to the MM part of the simulation. The electronic structure at the boundary will be very different if the hydrogen atom and the atom it replaces do not have similar electronegativities. These can be adjusted in semi-empirical methods, but can still cause issues especially for excited states. Further, most force fields do not include any concept of polarizability, but the QM region will redistribute charges.
For certain configurations, this can lead to significant imbalances and amplification of errors. The best practice to mitigate these potential problems is to choose boundary locations where there is minimal charge fluctuation expected, e.g. carbon atoms at least three unconjugated atoms away from any chemical reactions. Schrödinger’s Q-Site (111) program has many built-in scripts to assist with covalent embedding and implements frozen orbital methods as well. All covalent embedding schemes should be treated with ample caution as it is possible to break almost every implemented scheme. As always, rigorous testing with multiple simulations and perturbations of any QM/MM partitioning is a necessity.

Ultimately, it is still not clear if QM/MM actually helps to alleviate the problems of either QM or MM or just leads to more difficulty in and of itself.

### 1.3.6 Ligand Docking

A major area of research where virtual techniques are greatly sought is in drug discovery. Pharmaceutically active compounds work by binding to specific proteins and either activating or inhibiting them to some degree. The primary pedagogy in making a drug is to maximize how long it sticks to the intended target. For the most part, the longer, the better. To this end and in many diverse forms, molecular docking has been developed as a virtual way to quantify the number and relative strength of potential interactions of a molecule with a protein target of interest.

The more common docking procedure is rigid receptor docking, where the protein or receptor is frozen and held fixed in space and some conformational selection or energy minimization is performed on the ligands to generate an optimal pose in the binding site using a simplified grid map. A docking grid is a characterization of the binding site’s electrostatic
potential and key interactions in a 3D lattice. The program then estimates the free energy of
binding by implementing some form of energy-based scoring function that scales ($\alpha$ and $\beta$) and
sums some number approximated energetic contributions ($E$) like the generalized formula below.

$$Docking\ Score = \alpha E_{vdw} + \beta E_{Coulomb} + E_{hydrophobic} + E_{torsion} + E_{Hbond} + E_{metal} + \ldots$$  \hspace{1cm} (Eq. 14)

For rigid receptor classes of docking routines, the protein structures utilized must be in
the lowest energy, i.e. most probable, state. Crystal structures are a necessary starting point, but
due to substrate conformational selection or induced fit effects, significant computational studies
must be done to understand the structural dynamics of the protein. The crystallographic structure
is a position averaged snapshot of the lowest energy structure for that system. Often the
uncertainty in dynamic motion is needed for binding (conformational searching) and introducing
or excluding components will alter the system energetics (induced fit). Sufficient state sampling
with MD and/or MC is vital for proper state selection when beginning a docking simulation and
multiple crystal structures and conformations should be used to check convergence. Additionally,
few crystal structures possess atomic resolutions small enough to assume the protonation state of
a protein and often require non-physiological pHs in the crystallization methodology. Due to
these issues, special attention must be made to addressing side chain protonation states in
docking runs (see section 1.3.1.1 Determination of Protonation States).

Other docking schemes also exist to overcome the limitations of rigid receptor docking,
such as flexible or induced fit docking (Section 1.3.6.2 Induced Fit Docking) where the receptor
is essentially allowed to move and quantum polarized ligand docking that combines docking
with QM/MM techniques (Section 1.3.6.3 Quantum Polarized Ligand Docking).
For many decades, dozens of different docking programs, procedures, and associated tools have been developed for both commercial and academic use. This section will focus on the more popular programs Autodock Vina (112), DOCK (113), GLIDE (114), GOLD (115), and LigandFit (116) to give a general overview of what rigid receptor docking programs are available.

Table 5. List of docking software

<table>
<thead>
<tr>
<th>Program</th>
<th>Sampling Method</th>
<th>Scoring Method</th>
<th>Accuracy</th>
<th>License</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Scorea</td>
<td>Poseb</td>
</tr>
<tr>
<td>Autodock Vina</td>
<td>genetic algorithm</td>
<td>Force field</td>
<td>0.485</td>
<td>0.726</td>
</tr>
<tr>
<td>DOCK</td>
<td>shape-based</td>
<td>Force field or contact score</td>
<td>0.445</td>
<td>0.591</td>
</tr>
<tr>
<td>GLIDE</td>
<td>systematic search</td>
<td>Empirical score</td>
<td>0.544</td>
<td>0.754</td>
</tr>
<tr>
<td>GOLD</td>
<td>genetic algorithm</td>
<td>Empirical score</td>
<td>0.599</td>
<td>0.726</td>
</tr>
<tr>
<td>LigandFit</td>
<td>Monte Carlo</td>
<td>Empirical score</td>
<td>0.479</td>
<td>0.689</td>
</tr>
</tbody>
</table>

a Score accuracy is defined by the rate of the top scoring pose having less than 2Å RMSD from the native pose. b Pose accuracy is defined by the rate of a program being able to obtain a pose less than 2Å RMSD from the native pose from all pose outputs. (117)

AutoDock Vina was designed as a successor to the popular AutoDock developed by Trott and Olson (112) at the Scripps Research Institute in La Jolla, California. Like its predecessor, AutoDock Vina is a free, open source program and shares some basic approaches as the original AutoDock, using a genetic algorithm for processing poses and a MM based scoring function. However, many independent tests of AutoDock have shown it is no longer considered a sufficiently accurate docking software and has begun to fall out of use despite a large number of published articles owing to the free license (117,118). AutoDock Vina was developed to address
these issues and overhaul the program to offer significant improvements in the average accuracy of binding mode predictions, while also being up to two orders of magnitude faster than AutoDock 4 (119).

LigandFit is a CHARMM based program that uses MC techniques to sample conformations of ligands initially docked into an active site based on shape followed by further energy minimization. Originally developed by Venkatachalam et al. (116), LigandFit has scripts capable of identifying an active site and generating grids using a cavity detection algorithm. The program then fits a given ligand into the specified binding site through MC conformational sampling whose parameters depend on the number of rotatable bonds in docked molecules and matching the ligand to grid points. Those conformations then finally undergo energy minimization and scoring based on the DockScore energy function.

DOCK is docking software initially developed by Irwin Kuntz (120) at the University of California at San Francisco and versions have been available since 1988 free of charge for academic institutions. The most current version, DOCK 6 (113), utilizes several improved scoring options that include explicit terms for ligand and receptor desolvation and ligand conformational entropy corrections. A scoring function based on AMBER with implicit solvent, conjugate gradient energy minimization, and MD simulation capabilities are also present.

In 2004, the docking software GLIDE (Grid-based LIgand Docking with Energetics) was developed by Friesner et al. (114) and is currently a module in the Schrödinger software suite, commercially licensed by Schrödinger, LLC. GLIDE has become one of the more regarded docking programs with its relatively high pose and energetic accuracies (See Table 5). More recent docking software has shown even more promise in terms of speed and accuracy, but GLIDE has reliably held its position for many years. GLIDE works by generating a set of grids
with different types of fields representing the receptor binding site properties and geometries and then, unique to GLIDE, it performs a complete systematic search of the orientational, positional, and conformational space of the docked ligand. The best conformation is then further refined through MC sampling. Afterwards, the torsional space of the ligand is exhaustively sampled to generate a large number of potential binding poses. A hierarchical series of filters and scoring is employed following the initial rough positioning to narrow down the range of alternatives to be evaluated. A proprietary scoring function called GLIDEScore is used. Further minimization is applied leading to a final energy evaluation with a composite scoring function that combines empirical and MM force field based terms.

GOLD (Genetic Optimization for Ligand Docking) is another highly regarded, commercially available docking program created in a collaboration between the Cambridge Crystallographic Data Center (CCDC), GlaxoSmithKline, and the University of Sheffield (115). The program implements a genetic algorithm based search method for generating ligand poses and allows for some partial protein flexibility up to ten residues defined by the user. GOLD also deploys a user interface with an interactive docking setup called Hermes that contains a variety of constraint options and allows for the automatic consideration of cavity bound water molecules. Multiple proprietary and standard scoring functions can be considered including ChemScore, GoldScore, Piecewise Linear Potential (PLP), and the Astex Statistical Potential (ASP). This range of assets has kept GOLD as one of the better docking platforms available.

Schrödinger’s docking software was utilized in this work as the primary docking software due to the significant interconnectivity of programs, extensive literature review of the related algorithms, and prior access to licenses. The following subsections describe Schrödinger’s docking and energy scoring algorithms.
1.3.6.1 SP and XP Docking. SP (Standard Precision) docking is the default and original Glide scoring function (Eq. 15) using a series of hierarchical pose filters that perform conformational searches of a ligand set against shape and properties of a receptor grid. It remains one of the best docking algorithms according to benchmarking to date (Table 5) and provides a quick estimation of the free energy of binding based on extensive empirical validation. The Glide SP protocol has a reasonably fast docking speed (estimated 10-15 seconds/compound, overhead excluded), but suffers considerable slowdown for larger molecules with more rotatable bonds (114). XP (Extra Precision) docking is similar to SP, but uses a different energy function (Eq. 16) and sampling approach to better resolve hydrophobic interactions. The XP scoring functions accomplishes this by including water desolvation terms, ligand binding terms to specific protein structural motifs, and incorporating hydrophobic enclosure terms with modified hydrogen bonding conditions. This makes XP superior for interpreting binding sites with substantial hydrophobic characteristics and/or complex hydrogen bonding in hydrophobic enclosures. Unfortunately, this modified approach sacrifices substantial speed (approximately 2 minutes/compound) with the same dependency on number of rotatable bonds (121).

The main drawback for these docking programs are that both schemes are rigid docking methods and do not account for protein flexibility. Further, while Glide does attempt exhaustive pose screening, docking programs in general still suffer from poor ergodicity in possible pose prediction evident by a significant decrease in accuracy and speed for larger molecules (approx. >20-30 rotatable bonds). They also are dependent on normalizing arbitrary energy calculations with experimental values and still persist with relatively high error (averaging approximately +/- 2 kcal/mol). Appropriate renormalization with control compounds and proper protein preparation
can alleviate or at least codify some of this error, but the subjective nature of the scoring function will and should always yield suspicion to any docking results.

\[ GScore(\text{SP}) = 0.05 \left( E_{\text{vdw}} + 0.15 E_{\text{Coulomb}} + E_{\text{Lipo}} + E_{\text{Hbond}} + E_{\text{metal}} + E_{\text{rewards}} + E_{\text{Rot}} + E_{\text{site}} \right) \]  
(Eq. 15)

\[ XPScore = E_{\text{vdw}} + E_{\text{Coulomb}} + E_{\text{hyd_enclosure}} + E_{\text{hb_nn_motif}} + E_{\text{hb_cc_motif}} + E_{\text{pi}} + E_{\text{hb_pair}} + E_{\text{phobic_pair}} + E_{\text{desolv}} + E_{\text{ligand_strain}} \]  
(Eq. 16)

### 1.3.6.2 Induced Fit Docking

IFD (Induced Fit Docking) is one of the few docking algorithms capable of accounting for protein flexibility. IFD uses the docking program Glide to account for ligand flexibility, and the Refinement module in the Prime program to resolve receptor flexibility (16,122–125). The Schrödinger IFD protocol attempts to model induced-fit effects from alterations in binding site conformation due to ligand binding in order to increase the accuracy of binding affinity estimates and prediction of possible binding modes. IFD starts with initial ligand docking by Glide, employing a greatly reduced van der Waals radii and can temporarily remove highly flexible side chains during this docking step. Following primary rigid docking, the Prime structure prediction module deletes the residues surrounding the docked ligand and reintegrates them into the site while energy minimizing them. Lastly, IFD executes a final ligand re-docking to the recently generated low-energy protein structures and the resulting complexes are scored via another function unique to IFD (Eq. 17). IFD can also use either SP or XP scoring functionals for the GScore term, allowing proper definition of potential hydrophobic features.
The major disadvantage of IFD is the sizable computational time required for an IFD run, on the orders of hours per molecule. Further, handling entropy is still a major issue for most molecular mechanics. The IFD protocol does not account for entropy or work required to get to the calculated state. This can lead to very high estimated binding affinity for some theoretical poses where the energy barrier of the conformation represented would be physically impossible or extremely fleeting, such as hydrophobic ligands being buried in core residues. Because of this, it is advised to not use IFD as a general docking protocol, only implement IFD for flexible binding sites and rigid docking for inflexible ones.

\[
\text{IFDScore} = E_{\text{prime}} + 9.057 \text{GScore} + 1.428 E_{\text{Coulomb}(\text{GScore})}
\]  
(Eq. 17)

1.3.6.3 Quantum Polarized Ligand Docking. QPLD (Quantum Polarized Ligand Docking) is another Schrödinger docking program designed to compute exact energies through \textit{ab initio} quantum calculation and allows for the treatment of ligand polarization (126,127). While some advancements have been made in polarizable force fields (Section 1.2.4 More Complex Models), classical force fields in general are not capable of considering induced charge polarization. Docking methods rely on classical MM force fields and when charge polarization plays a larger role in the binding mechanism of a particular protein, the accuracy of the docking method is reduced. The QPLD protocol alleviates this limitation by generating multiple binding poses through conventional docking and then performs a single-point energy calculation on each complex and derives new partial atomic charges from electrostatic potential fitting. The updated molecule is then re-docked and scored with better handling of induced
polarization due to a highly charged active site or for a ligand that is particularly susceptible to polarization.

Similar to IFD, QPLD is very computationally expensive with simulations requiring hours per molecule. QPLD is also a rigid docking method, not accounting for protein flexibility in any way. Additionally, docking tautomers of the ligand is a rough, low-resolution, approach that has been proven to increase hit rates by approximating repolarization effects on atoms with a labile hydrogen (128,129). If an active site can induce and stabilize a tautomeric form of a molecule, a simple rigid docking of explicitly defined tautomers can simulate QPLD to some extent. QPLD is needed to explore more subtle effects of highly charged binding sites or of molecules that are susceptible to polarization without labile hydrogens.

\[
\Delta G_{QPLD} = E_{valence}(P) + E_{coul}(P) + E_{vdw}(P) + \alpha E_{vdw}(P/L) + E_{QM}(L) + G_{solv}(C) + G_{ideal}(C) \quad \text{(Eq. 18)}
\]

where P is the protein, L is the ligand, and C is the complex.

1.3.6.4 Constituent Partitioning Consensus Docking. Each docking algorithm previously described is fine tuned to accommodate and accurately analyze specific circumstances and conditions a ligand-protein complex may have. However, these programs are often not well-suited outside their intended function or calibration range. Further, while the specific cases for which these methods have been attuned are studied extensively, proper implementation practices are poorly defined. Consensus docking is the practice of combining more than one virtual screening program with differing scoring functions to improve screening results in a complementary fashion. However, all current approaches to consensus docking only utilize rigid
docking methods (130–136). Constituent Partitioning (CP) consensus docking attempts to apply
docking protocols that specialize in determining pose prediction and scoring from more
encompassing physical properties (e.g. polarizability, receptor flexibility, solvent interaction, etc.)
and weighting the output of each simulation to yield more insight into individual protein-ligand
complexes. CP consensus docking with appropriate controls helps to resolve the optimal
algorithm to use without needing direct measurements of the complex being simulated.

Table 6. Interpretations of Binding Site Characteristics from CP Consensus Docking.

<table>
<thead>
<tr>
<th>Docking Procedure</th>
<th>Docking Scores</th>
<th>Docking Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>High Correlation</td>
<td>Low Correlation</td>
</tr>
<tr>
<td>SP</td>
<td>- Rigid site backbone</td>
<td>- Flexible site backbone</td>
</tr>
<tr>
<td></td>
<td>- homogeneous charge distribution</td>
<td>- heterogeneous charge distribution</td>
</tr>
<tr>
<td></td>
<td>- Minimal hydrophobic and solvent interaction</td>
<td>- Significant hydrophobic and solvent interaction</td>
</tr>
<tr>
<td>XP</td>
<td>- Significant hydrophobic environment with hydrogen-bonding residues</td>
<td>- Minimal site hydrophobicity</td>
</tr>
<tr>
<td>IFD</td>
<td>- Induce fit upon ligand binding</td>
<td>- No induced fit effects upon binding</td>
</tr>
<tr>
<td></td>
<td>- Flexible backbone site</td>
<td>- Rigid site backbone</td>
</tr>
<tr>
<td>QPLD</td>
<td>- Highly charged site</td>
<td>- Homogeneous charge distribution</td>
</tr>
<tr>
<td></td>
<td>- Docked ligands highly susceptible to repolarization</td>
<td>- Docked ligands are not polarizable</td>
</tr>
<tr>
<td>MM/GBSA</td>
<td>- Significant solvent interactions</td>
<td>- Solvent interaction minimal</td>
</tr>
<tr>
<td></td>
<td>- Highly flexible site backbone</td>
<td>- Low site flexibility</td>
</tr>
</tbody>
</table>

The first task is to identify control molecules with accurate, experimentally derived
binding affinities. From practice, approximately 8 – 12 compounds are sufficient for generating
linear regression statistics with the current degree of docking error, connoting a diminishing
return for larger numbers of controls. Additionally, a wide range of binding affinities is required
(strong binders, moderate binders, and non-binders as decoys). After run each docking protocol, a Pearson sample correlation function (Eq. 19) is applied to estimated free energies of binding ($\Delta^\circ G$) from the docking runs and the known experimental binding energies (Table 15 gives an archetypal example of this).

$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \langle x \rangle)(y_i - \langle y \rangle)}{\sqrt{\sum_{i=1}^{n} (x_i - \langle x \rangle)^2} \sqrt{\sum_{i=1}^{n} (y_i - \langle y \rangle)^2}}$$

(Eq. 19)

Where $x_i$ are the individual docking scores for each control compound, $y_i$ are the associated experimentally derived energies, and $\langle x \rangle, \langle y \rangle$ are the sample averages. If the experimental binding affinities are given as a $K_D$, they must be converted to $\Delta^\circ G$ (Eq. 20) as Eq. 19 is meant for linear correlations.

$$\Delta^\circ G = RT \ln (K_D)$$

(Eq. 20)

Where $R$ is the ideal gas constant and $T$ is the temperature at which the experiments were run. $r$ has a value between -1 and +1 and the strength of correlation depends on experiment and number of values used for correlation. For the following studies, a basic guide was developed where $r > 0.9$ was considered a very strong correlation, $0.9 > r > 0.8$ was strong, $0.8 > r > 0.7$ was moderate, and $r < 0.7$ was deemed poor. The strength of correlation helps to infer protein dynamics and binding characterizations needed to understand binding mechanisms, build better
computational models, and increase accuracy and reliability of data from modeling by being able to delineate sources of model error.

Model results with the higher Pearson correlation for controls are the optimal docking protocols for that particular protein (Table 6). Strong correlations also imply certain postulates of dominant binding site features. Proteins with rigid backbones, homogeneous charge distribution, minimal hydrophobic interactions, and low potential for solvent intrusion in the binding site are likely best described by SP docking (Section 1.3.6.1 SP and XP Docking). The same binding site with a greater degree of hydrophobic enclosure would be better interpreted by XP docking. More flexible sites require IFD (Section 1.3.6.2 Induced Fit Docking) and can still be simulated under SP or XP scoring functions. Higher correlations under QPLD (Section 1.3.6.3 Quantum Polarized Ligand Docking) demonstrate dependence on ligand repolarization upon binding. These sites have more extreme charge distributions and will induce polarization on a susceptible ligand. More non-polar ligands may not be as sensitive to electrostatically heterogeneous environments, so some chemoinformatic metrics on the ligands being docked are important to exclude the need for the higher computational cost of QPLD. MM/GBSA (Section 1.3.2.3 MM/GBSA) coupled with short MD runs is ideal for highly flexible binding sites with a significant degree of possible solvent interaction since the Generalized Born and Surface Area continuum solvation accounts for solvent effects without the need for explicit solvent molecules over a trajectory. These methods combine to give a more detailed understanding of a protein binding site. For the computational cost of a few dozen docking simulations, a commensurate representative algorithm can also be deduced, helping to discard unnecessary calculations (See Table 15 for an example). If a simplified rigid docking protocol is sufficient to predict binders, then any further calculation is wasted effort. CP consensus docking improves the overall
accuracy and performance of individual models, reduces the computational burden for drug discovery projects, and provides unique mechanistic insight into binding dynamics for any given protein.

1.3.7 Virtual Target Screening

Virtual Target Screening (VTS) is a system designed to virtually screen a molecule of interest (MOI) to a large library of protein structures. The current protein library consists of 1,451 structures. The system is calibrated with a set of small drug-like molecules that are docked against each structure in the protein library to produce benchmark statistics. VTS was originally employed as a theoretical assay of potential kinase activity and gauge of potential biological promiscuity. The calibration procedure allows the analysis to accurately predict inhibitor–kinase binding affinities when $K_D < 10 \ \mu M$ (defining a hit) and $K_D \geq 10 \ \mu M$ are both considered (72% accuracy in the best case) (137). Therefore, the VTS system is able to robustly discriminate protein binders from nonbinders and give some inclination as to potential binding promiscuity of the molecule of interest with respect to the protein group tested.

1.3.7.1 Calibration. As stated previously, scoring algorithms are defined in a fairly arbitrary manner. The ligand-protein complexes used as training sets to scale the output metrics of these scoring functions carry over bias and are not diverse enough to span the continuum of possible interactions. Four major echelons of increasing difficulty must be reached for a molecule to be considered a hit:

1) Overcome the intrinsic thermal energy of water.
2) Out-compete other random molecules for the binding site.

3) Displace the native binding partners of the binding site in question.

4) Optimize specificity for the binding site.

Additionally, for drug discovery at least, the affinity for which a ligand binds to a receptor is not the most useful number. One of the more important characteristics that makes a good drug is its ability to displace the natural ligand from the receptor, or apparent inhibition. If a molecule has an impressive picomolar binding affinity, the compound’s relevance is nearly null if the native ligand binds in the femtomolar regime (e.g. (stept)avidin-biotin or Ras GTP binding). Conversely, an estimated $K_D$ of single digit micromolar is relatively substantial compared to a biological effector affinity near the millimolar range (e.g. Neurotransmitters or many kinase’s affinity for ATP). Further, if a binding site has a more generalized range of binding partners (e.g. esterases and lactamases), the binding sites are adapted to be promiscuous for more random types of molecules. Docking algorithms are unable to account for these general cases that comprise a significant degree of protein binding modes. Therefore, to inter-compare docking scores between disparate protein binding sites, a common metric must be developed to properly renormalize the arbitrary scoring into a physically definable characteristic for each individual protein binding site.

Our approach utilizes a diverse compound set to represent a random molecule binding to a site and analyzing the ability of that site to discriminate between random compounds and known binding ligands. Then, employing robust statistical analyses, create a universal ruler with which to measure each protein binding site’s propensity to bind a random molecule. This technique gives a lower bound on what a “good” docking score is for any particular binding site.
There are certain caveats to this and are expounded on in the following section (1.3.7.2 Scoring and Confidence Intervals). To further adjust docking scores toward differentiating between moderate binders and potent inhibitors, the native ligand or known binder can be used as a control to renormalize and test if the algorithm is capable of yielding physically relevant results. Coupling these two factors with Eq. 22 and the ligand specificity metric (Section 1.3.7.3 Ligand Specificity) uniquely solves each separate issue mentioned above.

For calibration, each protein must be adequately prepared (Section 1.1 Preparation and Best Practices for Modeling) and identify each possible binding site. A diverse set of molecules is then docked to the site. The original prototype (137) used the NCI Diversity Set (138) for this purpose. However, statistical analysis (Figure 1) showed poor diversity of the NCI diversity set and led to developing a new calibration set based on a theoretical compilation of billions of virtual chemical entities designed by Virshup et al (139). This small molecule universe (SMU) set is superior since the compounds are only meant to probe the virtual site and should not biased toward any existing chemical entity. High diversity of this compound list is also required to properly sample as much of the possible chemical space within practical bounds. Figure 1 also shows the far greater chemical diversity of the SMU set, prompting its use in current iterations of VTS.
Figure 1. Diversity analysis of SMU (Left) and NCI diversity set (Right). A Self organizing map of 116 weighted chemical descriptors was computed. Each region is colored from red to blue indicating the number of compounds in each sector. The number of compounds define in each sector is also shown in the colored box. The distance between each mapped value is represented by gray-scale connecting lines between the boxes with the scale legend on the bottom.

Calibration gives a relative lower bound on what docking score values should be considered as a hit. However, further refinement may be needed for validation of binding metrics beside estimated free energy of binding. The ability of a docking program to reproduce crystallographic binding poses within thermal limits is also necessary. As such, pose evaluation is also automatically performed if a co-crystallized ligand is detected. A 2Å average deviation is expected from thermal fluctuations (Table 3) and is the commonly accepted threshold for evaluating docking poses (140). A “self-dock” procedure is utilized where the co-crystallized ligand is evaluated by the docking program. The following section (Section 1.3.7.2 Scoring and Confidence Intervals) details the incorporation of self-dock metrics into the final VTS scoring.
If reproducible literature values also exist for the control ligands, that information is evaluated as well. However, the ongoing reproducibility crisis (141–145) is showing affect in drug discovery as many biochemical assays performed throughout literature are consistently found to be dubious and incorrect (146,147). Between the poorly understood severity of molecular aggregation(148–151), limited standardization of assay protocols (152–155), and co-equal treatment of surrogate species for human systems (156–158), most databases on biochemical assay results are heavily contaminated with these unreliable results and are, therefore, useless as a control. This issue has severely limited the use of statistically significant sample sizes for assessing the physical application of binding energy estimations.

Following calibration and native ligand control analysis, the MOI is docked to each processed protein in the library and the scores are related to the calibration metrics for final scoring.

1.3.7.2 Scoring and Confidence Intervals. The original VTS prototype (137) relied on qualitative measurement to relate the VTS calibration to the MOI. A comprehensive scoring algorithm (Eq. 21) was created to evaluate MOI docking runs to protein grid database calibration metrics in a quantitative and statistically robust manner. The functional is based on a weighted average of confidence interval calculations derived from distribution regression analysis.

\[
VTS_{\text{Score}} = F(x_{d,\text{score}}) \left[ \frac{(VTS_{\text{SD}})(SDME)(SD_{\text{CI}})+(VTS_{\text{CDF}})(VTS_{\text{CI}})(VTS_{\text{Boltz}})}{(SDME)(SD_{\text{CI}})+(VTS_{\text{CI}})(VTS_{\text{Boltz}})} \right] \quad \text{(Eq. 21)}
\]
The score is dichotomized into two primary weights, native ligand self-dock comparison and VTS comparison, where each functional term is defined as follows (Eqs. 22-28),

\[
F(x_{d\_score}) = 1 - \frac{1}{\pi} \arctan \left( \frac{x_{d\_score} - C_{H,O}}{\sqrt{k_B T}} \right) + \frac{1}{2} \quad \text{(Eq. 22)}
\]

With \( x_{d\_score} \) as the MOI docking score, \( C_{H,O} \) as the heat capacity of bulk water (5.611 kcal/mol for 310K body temperature, 5.410 kcal/mol for 298K room temperature), \( k_B \) as the boltzmann constant, and \( T \) as temperature. \( F(x_{d\_score}) \) is a function intended to give a confidence interval formulation for quantifying probability of binding when compared to the inherent thermal energy distribution in bulk water. The expected distributions of MOI docking scores are estimated as a Gaussian distribution with a documented error (114) of +/- 2 kcal/mol average docking energy error taken as the standard deviation (\( \sigma_{\text{Glide}} \) in Eq. 24). The thermal energy distribution in bulk water is estimated as another Gaussian with the thermodynamic resolution limit of \( k_B T \) as the expected variance. Taking these two Gaussians as a ratio yields the probability distribution of the MOI-protein complex remaining associated. This distribution is also represented as a Cauchy distribution. Integrating the Cauchy distribution for the given MOI docking score and heat capacity allows for a confidence interval to be calculated from the density function (Eq. 22).

The self-dock docking score (SD\_score) is the docking score from re-docking a co-crystallized ligand back into the binding site. This allows for control over normalizing both the calibration set and docking algorithm performance with respect to a known binder. The Self-
Dock Confidence Interval (SD_CI, Eq. 23) is the normal cumulative distribution function giving
the confidence interval of the self-dock ligand having a better (In this case, a more negative
estimated binding energy) expected docking score than the calibration set.

\[
SD_{\text{CI}} = 1 - \int_{-\infty}^{\text{SD_Score}} \frac{1}{\sigma_{p_{\text{cal}}_{\text{total}}} \sqrt{2\pi}} e^{\frac{(\tau - \mu_{p_{\text{cal}}_{\text{total}}})^2}{2\sigma_{p_{\text{cal}}_{\text{total}}}^2}} d\tau
\]  

(Eq. 23)

Where \( \mu_{p_{\text{cal}}_{\text{total}}} \) and \( \sigma_{p_{\text{cal}}_{\text{total}}} \) are the average and standard deviation of calibration set
docking scores for a particular target, respectively. Both of these values are derived from Figure
2 (mu = \( \mu_{p_{\text{cal}}_{\text{total}}} \), Sigma = \( \sigma_{p_{\text{cal}}_{\text{total}}} \)). If the SD_CI is insignificant, the docking
algorithm is not adequate for reproducing proper controls for that particular protein-ligand
complex and the weighted average will only rely only on the calibration set.

For defining a probability that a MOI docking score lies outside the calibration
expectation range, the average calibration set docking scores for each individual target protein
binding site are averaged and fitted to a normal distribution (Figure 2 and Eq. 24).

\[
VTS_{\text{SD}} = 1 - \int_{-\infty}^{x_{\text{min}}} \frac{1}{\sigma_{\text{Glide}} \sqrt{2\pi}} e^{\frac{(\tau - \text{SD_Score})^2}{2\sigma_{\text{Glide}}^2}} d\tau
\]  

(Eq. 24)

The VTS Self-Dock comparison score (VTS_SD) sets a lower bound on the minimum
docking score for a MOI to be statistically relevant. An interesting note is that the average
calibration docking score average (mu in Figure 2) is slightly lower than the average thermal
energy of water (~5.5 kcal/mol) and the standard deviation (Sigma in Figure 2) is near \( k_B T \)
(~0.593 kcal/mol at 298K). This may imply some emergent evolutionary cutoff for protein binding sites to assist in selection of the intended native ligand, using the inherent thermal energy of water to prevent non-specific molecules from binding.

Figure 2. Gaussian distribution fit of calibration set average docking scores.

![Gaussian distribution fit](image)

Standard deviation of calibration set docking scores on a target protein binding site can be used to evaluate the expected error intrinsic to the binding site (Eq. 25).

$$VTS_{CI} = 1 - \int_{-\infty}^{\infty} \frac{1}{\tau \sigma_{\text{cal, total}}} \sqrt{2\pi} e^{-\frac{(\ln \tau - \mu_{\text{cal, total}})^2}{2\sigma_{\text{cal, total}}^2}} \, d\tau$$  \hspace{1cm} \text{(Eq. 25)}$$
Where $\sigma_{p_{\text{cal, total}}}$ is the standard deviation of standard deviations. This variance analysis is normally handled by a $\chi^2$ distribution (159). However, the probability density function needed for a confidence interval of a $\chi^2$ distribution has no simple forms. Therefore, estimation with a log-normal distribution is preferred and found to be an acceptable approximation (160). The VTS Confidence Interval (VTS_CI, Eq. 25) is then calculated from fitting the calibration set docking score standard deviation to a log-normal function (Figure 3), allowing for estimation of expectation values, or likely range of scores, for the protein binding site. If the standard deviation of calibration set docking scores is too high compared to average values (mu in Figure 3), then the VTS algorithm is not capable of discriminating between random molecules, control ligands, and MOIs due to high variance in the associated docking scores. This term drops the relevance of calibration set comparison in the final scoring (Eq. 21).

The Self Dock Mean Error (SDME, Eq. 26) is used to calculate the difference between the native ligand re-dock pose and the crystallographic pose and relate it to a confidence interval.

$$SDME = \left( \frac{\text{RMSD}}{\text{RMSD}_{\text{min}}} \right)^{-1.363}$$

(Eq. 26)

Where RMSD is the calculated root mean squared deviation of the native ligand re-dock pose from the crystallographic pose and RMSD_{min} is the minimum cutoff for evaluating docking poses estimated at the commonly accepted threshold of 2Å (140). The exponent is determined from a power-law distribution fit of Self-Dock RMSD values (Figure 4). Eq. 26 is a simplified approximation for obtaining a probability distribution of the associated power law (161). For obvious reasons, if the SDME is insignificant, then the docking algorithm is not able to
reliability reproduce physical data and results in the weighted average, VTS_Score (Eq. 21), to rely only on calibration data.

![Log-normal distribution fit of calibration set standard deviations](image)

**Figure 3.** Log-normal distribution fit of calibration set standard deviations.

Testing if the calibration set top score distribution has better docking energies than bulk water helps to assess if the VTS scoring method is even relevant to the target protein binding site.

\[
VTS_{\text{Boltz}} = 1 - \frac{1}{\pi} \arctan \left( \frac{x_{2.5\sigma} - C_{H,O}}{\sqrt{k_B T}} \right) + \frac{1}{2}
\]  

(Eq. 27)

**Eq. 27** is the same form of **Eq. 22** with \(x_{2.5\sigma}\) as the top 1.24% of calibration molecule docking scores. Top scores are defined by the most negative docking energies. The percentage was made
to closely approximate the top 20 and top 200 metrics and Boltzmann factoring (VTS_Boltz) used in the original VTS prototype (137). If the VTS_Boltz score is too low, the the weighted average, VTS_Score (Eq. 21), will rely primarily on self-docking.

Figure 4. Power law distribution fit of self dock RMSD.

The primary VTS metric is comparing the MOI docking score to calibration set average docking score. This is done by integrating over the normal distribution of calibration scores evaluated at $x_{d\_score}$, which is the MOI docking score. This produces a confidence interval from the cumulative distribution function of MOI docking scores compared against the calibration set (VTS_CDF, Eq. 28).

$$
VTS\_CDF(x_{d\_score}) = 1 - \int_{-\infty}^{x_{d\_score}} \frac{1}{\sigma_{p\_cal\_total} \sqrt{2\pi}} e^{\frac{(\tau - \mu_{p\_cal\_total})^2}{2\sigma_{p\_cal\_total}^2}} d\tau
$$

(28)
VTS_Score is then defined as the certainty of a “hit” with scores between 0, defining a model with absolutely no predictive value, and 1, perfect predictive model fit. The confidence interval method is used to provide an intuitive measurement of the likelihood a MOI will bind to a target protein with an arbitrary cutoff of estimated binding affinity. The cutoff used for this work is 10 μM $K_D$ or -8.505 kcal/mol at 310K (Eq. 20). A VTS_Score of 0.95 will correspond with the estimated binding affinity required to define a hit for a particular target being -8.505 kcal/mol at 310K, indicating an approximately 95% chance that the MOI will bind with a 10 μM $K_D$ or less. This corresponds closely with a prediction interval of 2σ. Increasing the cutoff can help decrease the false positive rate while sacrificing true positives.

1.3.7.3 Ligand Specificity. The VTS protocol has many diverse applications in biochemistry and drug discovery efforts. One of particular promise is the possibility for assessing a ligand’s effectiveness to exclude binding to diverse protein sites. This Ligand Specificity (LS) analysis can be used to infer the promiscuity of a MOI and related compounds. Even with a limited subset of proteins in the prototype library, the capacity for a ligand to differentiate binding pockets can be quantified. Fitting the final VTS score of a MOI from each protein to a beta distribution (Eq. 29) allows the estimation of the mean (Eq. 30) VTS_Score for that MOI.

\[
f(x; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}
\]

Where $\Gamma$ is the gamma function, and $\alpha$ and $\beta$ are the shape parameters to be fitted. The expectation value, $E[X]$ or LS, is defined as follows,
Since the VTS scores are between 0 and 1, the expectation values for the beta distribution are also between 0 and 1. 0 is defined as a ligand being perfectly specific and 1 is defined as perfectly promiscuous, or binding all proteins with equal affinity. This measure can give significant information about potential off-target effects and understanding about the degree of biological selectivity for certain chemical compositions. Unfortunately, testing this parameter has been hampered by poor accounting of aggregation-based promiscuity in most chemical databases (148–151). The ligand specificity described here pertains to potential inherent non-specific binding motifs in chemical space, as opposed to colloidal aggregation. As such, obtaining a significant number of high quality binding data against a large number of biochemical targets controlled for aggregation to test the LS metric has proven untenable so far.

1.3.7.4 Enzyme Selectivity. The ability for an enzyme to select a single specific molecular substrate amidst a sea of similar chemical entities is often considered absolute. However, many enzymes have been known for sometime to bind and catalyze many different substrates and reactions. Some early examples of non-selective enzymes are chymotrypsin (162), pepsin (163), L-asparaginase (164), pyruvate decarboxylase (165), and carbonic anhydrase (166). Continued testing and characterization of proteins has slowly started to erode the old notions that protein binding sites are optimized through billions of years of evolution to recognize one, and only one, substrate (167). Our work on VTS has also shed unexpected light on this phenomenon of enzyme promiscuity toward substrate binding. Since the process of binding site calibration essentially exposes the site to a wide degree of chemical scaffolds, it is useful in characterizing
the site’s selectivity. The variance calculated from **Eq. 25** yields an expected range of scores for all protein binding sites in the database. **Figure 3** shows that most proteins are capable of binding only a small range of molecules, though many appear to be highly promiscuous. A simple fit to a normal distribution for each site generates an expected range of scores through estimation of standard deviation (**Figure 5**).

A wider distribution, or larger standard deviation, of docking scores indicates that a binding site can interact strongly with a broad range of chemical space, implying a limited ability for a site to discriminate between a variety of molecules. A thinner distribution with a small standard deviation demonstrates the site’s ability to discriminate against random molecular topologies, relating a higher selectivity for a specific molecular entity. This metric enables VTS to assess the target’s druggability. If a target is inherently promiscuous, it is likely not amenable to specific drug targeting and is, therefore, a weak candidate for drug discovery efforts.

### 1.3.8 Homology Modeling and Protein Folding

Homology modeling refers to regenerating the three-dimensional structure of a protein from the constituent amino acid sequence using known experimental structures of homologous proteins. Also known as comparative modeling or template-based modeling, homology modeling can be helpful when no other structural data exists as this structural information is vital in the study of protein function, dynamics, and understanding interactions with ligands and other proteins. The low-resolution structure predicted by homology modeling can contain useful information about spatial arrangements of important residues in a protein and may help guide the design of new experiments or structure identification, such as site directed mutagenesis.
Figure 5. Representative normal distribution fits of calibration set docking energies. Example of a highly promiscuous binding site (Top) and a relatively selective one (Bottom). Distribution fits marked as red line.
Obtaining structural data is not a trivial task and can be prohibitively difficult to solve experimentally, even taking decades to solve the structure of one protein. This process can be hindered by the difficulty of obtaining enough material through cloning, expression, and purification of milligram quantities of the protein, and difficulties associated with crystallization, such as post-translational side chain modifications and optimizing salt, pH, and temperature conditions. As it stands, genetic sequence determination far exceeds experimental structure determination and this gap will likely only grow. This presents a demand for faster computationally based approaches, though, substantial research in this area has yielded limited progress.

Current research and modeling techniques of *ab initio* quantum mechanical or classical MD simulations studying the prediction of protein structure from sequence alone consistently yields poor results (168). For QM, the computing power required to simulate even the smallest proteins is often untenable and must neglect environmental factors such as protein chaperons, membrane interaction, interactions with ribosome, and other numerous folding factors. While MD is computationally cheaper than QM, folding timescales are much longer than most feasible simulation timescales (see Table 3) for proper statistical sampling. Furthermore, current MM force fields are not sufficiently accurate for reproducing known folds; even lower and mid level QM are inadequate to the task. Chemical bonds and metal coordination are also difficult to mimic in a simulation, but are obviously required in real folding processes. Finding simpler physics-based rules that predict protein structure by looking at patterns of hydrophobic, hydrophilic, or charged amino acids has also been attempted for some time, albeit with little success (169,170).
Out of this disappointing imbroglio, some basic knowledge-based methods have emerged using a felicitous quirk in evolution that proteins with similar sequences tend to be homologs, evolving from a common ancestor (171). The fold of the protein also tends to be conserved during evolution; even some proteins with only 15% sequence identity can have similar structures. To this end, well over 100,000 protein structures have already been solved and that information could be used to help predict new structures. This is where template-based modeling (homology modeling) is used when one or more likely homologs of known structure can be identified. *Ab initio* structure prediction, or “*de novo*” prediction, is still needed if no homologs exist, but these approaches also take advantage of available structural data in subtler ways.

The sequence identity to the template structure and the quality of the sequence alignment are the major factors impacting the quality of a homology model. Gaps in alignment, commonly called indels, are an extensive complication that indicate a structural region present in the target does not appear in the template. Structure gaps in the template can also occur, caused by poor resolution in the experimental structure or when a deletion or insertion mutation exists in a solved structure introducing serious local errors (172). The use of multiple templates, referred to as composite modeling, can help alleviate some of the issues arising from gaps, but the differing local structures of the template around the gap and high probability of a missing region in one template also being absent from other structures in the same protein family complicate this method (173). Gaps are most common in loops where the difficulty of experimentally resolving the region greatly increases due to high local flexibility. Although some guidance can be provided from other templates, the longer the gap, the less accurate the model. Smaller gaps can be modeled with reasonable accuracy provided the local alignment is correct. Gaps beyond 8 or 9 residues require *ab initio* structure prediction techniques which do not often increase model
accuracy enough to afford the computational cost (174). The quality of a model also declines rapidly with decreasing sequence identity. At around 70-80% sequence identity, the RMSD between matching Cα atoms for a typical target is about 1-2Å, enough to start reversing dihedrals and impacting loop structures (See Table 3). This is still within normal crystal structure resolution. Yet in loop regions where the amino acid sequences of the target and template proteins may be completely different, errors are significantly higher. At 25% sequence identity, there is only a 3-4Å agreement at best (175). This is where crystal structures become unreliable as positional data accuracy for core domain residues deteriorates and information on loop structures is completely lost (see section 1.1.2 X-Ray Crystal Structure Resolution).

All said, homology models are useful for initial alpha carbon placement of key residues in folded proteins, assisting mutagenesis experiments, and forming hypotheses on structure-function relationships. Further, there has been significant progress in structure prediction over the last few decades due to improved algorithms and increased coverage in experimentally determined structural databases. Some of the most success is seen in GPCR receptor pharmacology and analysis of other transmembrane proteins that are notoriously difficult to resolve through experiment (176).

Unfortunately, even the latest accuracy benchmarks fall far short of experimental standards. The Critical Assessment of protein Structure Prediction (CASP) is a bi-annual, community-based competition for predicting protein structures that has been held since 1994 (177). CASP uses a Global Distance Test (GDT) metric that is analogous to RMSD, but isn’t influenced as much by outliers (178). As such, GDT tends to be much more forgiving than RMSD resolution standards. According to the latest CASP review (179), the most accurate homology models still struggle to achieve over 80% correct alignment on average for even the
simplest structures. For CASP, correct alignment of a modeled residue is when the predicted side chain’s Cα is less than 3.8Å from the corresponding experimentally derived residue position after optimal superpositioning. This is comparable to practical experimental cutoffs of 3-4Å described previously.

Homology models to date are ultimately unreliable in predicting conformational shifts induced by insertions or deletions or proffering any details of side-chain orientation. This issue discourages the use of homology models in ligand docking or drug design unless the sequence identity with the template is greater than 80% with at least 90-95% homology in the binding site. Even at this point, homology models are still considerably less reliable than an empirical crystallographic, NMR, or other experimentally derived structure. Moreover, at very high sequence identities, >95%, as many as one in ten models have an RMSD greater than 5Å compared to the empirical structure (20). Therefore, while the accuracy of predicted structures is increasing slowly but steadily, the issues associated with accurate refinement methods, sequence alignment, and template identification, still prevent a wider embrace of homology models in general research. Despite current difficulties, structure prediction is and will be even more critical in the future, demanding continued research by dedicated experts.

1.4 Informatics

Informatics involves the practice of information processing and the engineering of information systems to transform data, information, and knowledge into more useful or readable applications. Informatics and its various sub-disciplines are worth a book on their own and this section is only a very simplified discourse denoting some aspects relevant for the audience or referral to other chapters.
The most basic form of informatics involves storing and categorizing data. There exist numerous and wide ranging databases for biochemistry from the RCSB PDB repository for protein structures to PubChem (180,181) and ChEMBL (182–184) that serve as databases for chemical compounds and their associated biochemical activities in various assays. The next tier of informatics is to use this data by translating the language of chemical and biological results into a form machines can understand by way of machine learning, pattern recognition, regression analysis, and so on for designing and refining models widely applied in science, business, and government. This data mining can change the point of view of complex interactions, fortifying the quality of knowledge in any field and expanding its scope into others. Subsections of discipline specialized informatics include cheminformatics and bioinformatics that are highly beneficial to biochemistry.

1.4.1 Cheminformatics

Cheminformatics attempts to reduce the dimensionality of chemical space through characterization and categorization of molecular properties, pharmacokinetics, structure activity relationships, and pharmacophore modeling. Millions of compounds can be screened in seconds through substructure matching and fingerprint-based similarity searches with clustering and diversity selection to improve real and virtual compound libraries. Supervised learning techniques coupled with self-organizing maps that help reduce complexity, along with high dimensional information and principal components analysis, can produce quantitative models for structure-activity relationships and provide insight into new compound formulations. Programs that can computationally predict properties such as pKa, water solubility, intestinal absorption,
plasma-protein binding, etc. directly related to the biological effect of a drug and its subsequent metabolism, referred to as ADME-tox, are greatly valued in medicinal chemistry.

Canvas (185–187) is a user interface specific to chemoinformatic processing developed by Schrödinger. Chemistry Development Kit (CDK) (188,189) is an open source software alternative developed with the Java programming language for bioinformatics and chemoinformatics applications. These applications can computationally assay ADME-tox, yield chemical similarity metrics, and build pharmacophores and structure activity relationships (SAR) from experimental data, as well as many other precise functions. SAR prediction assumes that similar molecules will have similarity activities, therefore, once a standard curve is perfected from empirical data, only chemical similarity fingerprints are needed to predict binding affinity. Quantitative structure-activity relationships (QSAR) go another step further in addressing the supposed connection between chemical structure and biological activity by including additional theoretical descriptors of physio-chemical properties (163). While these metrics are invaluable, computational efforts in accurately assessing ADME have been fraught with difficulty (191,192). In addition to requiring substantial amount of rigorous data to build reliable models, ADME and SAR prediction software algorithms are minimally based in physical law and are and are notorious for misleading end-users with contradicting results. Overall, in silico approaches still largely lack the required predictability to fully engage their use in drug discovery (193).

1.4.1.1 Pharmacophore Modeling. Pharmacophore modeling utilizes a course-grained three-dimensional mapping of electronic and steric aspects of a biological target to inform and rank the fundamental molecular interactions required for ligand binding. A pharmacophore model attempts to reduce the essential features of the receptor or SAR down into simpler sub-
space definitions such as hydrophobic centroids, aromatic rings, hydrogen bond acceptors or donors, cations, and anions (194). Robust models can screen libraries consisting of millions of small compounds within hours or even minutes with varying precision. The molecules in these screening libraries are preprocessed to be in the lowest energy, bio-relevant form. This does impose some additional bias to the procedure, trading accuracy for speed (195). Each conformation is then fitted to the pharmacophore query by attempting to align the predetermined pharmacophore features of the target to the molecule. The pharmacophore query uses spheres to estimate optimal influence. If the essential features of a molecule fit in the spheres, the molecule is considered a hit. Partial matching can also be implemented if the query is found to be excessively complex for matching a given library by defining fewer features deemed more critical for binding.

Depending on what information is used to generate the model, ligand-based or structure-based pharmacophore are two methods for generating a model. The Schrödinger program Phase (196–198) can used to implement either strategy. If no empirical structure exists for a protein target, a ligand-based pharmacophore model is an indispensable tool for a drug discovery campaign. This strategy searches for the common chemical characteristics between many ligands with conclusive experimental binding data to extract the principal interactions requisite for binding to the target. Conversely, structure-based methods rely only on mapping critical binding features from the protein binding site directly. This type of model probes the site through various means, including fragment based docking and steric mapping, to generate a pharmacophore query. Despite the low resolution of the method, the pharmacophore approach is useful in rapidly generating potential hits to be followed up with more rigorous modeling and experiments.
1.4.2 Bioinformatics

Bioinformatics assists in combing genome, proteome, and metabolic data with specially designed methods and software tools for analysis and interpretation of immense biological databases. An inherently interdisciplinary field, bioinformatics incorporates mathematics and statistics with biology and computer science. There are many disparate bioinformatics software that focus on different divisions of cellular function.

For genome centric applications, sequence and structure analysis tools like EMBOSS (199) and clustalw (200) are useful for genetic sequence alignment, rapid database searching, and motif and domain analysis.

On the proteome level, many . With respect to structural information, the Protein Data Bank (PDB) is the definitive worldwide repository of large biological molecule 3D structures, including nucleic acids and proteins (7,8). For protein sequence alignment and database searching, FASTA (201) is the most widely used format in bioinformatics, using local alignments to a collection of fragments for similarity searches. BLAST (202–205) (Basic Local Alignment Search Tool) is another ubiquitous searching tool for perform fast similarity searches of proteins or DNA.

The domain of metabolic space encompasses all of the aforementioned branches, rendering an extreme complexity of subject area. As such, few programs are capable of spanning this discipline with any significant effect. Metacyc (206) is a curated database dedicated to studying metabolic pathways and the associated enzymes and metabolites and includes tools for database searching and limited pathway perturbation.
1.4.3 Best Practices in Statistics

Chemoinformatics and bioinformatics are relatively new fields referring to a novel branch of science straddling the traditional domains of biology, chemistry, and informatics. These discipline specific informatics attempt to exploit chemical and biological data by the creation and application of information-based methodologies. Therefore, the effective use of informatics in these domains necessitates a sound mastery of their underlying mathematical and statistical principles. First, an understanding of what is being measured is needed. Populations are difficult to measure in their entirely, so often smaller subsets are sampled to simplify experiments and are used to extrapolate details about the whole population. Separate and distinct quantities arise to describe the population itself versus the quality of its sample. The most common methods applied are the mean, median, and mode of a sample. These three quantities help approximate the values expected of the dataset distribution, but are not in any way interchangeable except in unique distributions rarely seen in any dataset and great care must be taken to understand which quantity should be used for minimal ambiguity. Other statistics characterize the spread in values of the dataset elements. Standard deviation is the square root of the variance, or mean squared deviation from the sample mean, and is the more useful descriptive quantities for understanding spread in a population dataset. Standard error, standard deviation divided by the number of samples, is only appropriate when discussing the quality of a sample (i.e. the expected deviation of a sample’s mean from the population’s true mean) and says nothing useful about the population distribution. Standard error will always be less than the standard deviation and is a common misleading statistical abuse to show smaller experimental error or significance. Standard error discusses error in a single experiment, whereas standard deviation discusses error across multiple independent experiments.
Statistics are intended to make data easier to understand but are often used in a misleading fashion in an attempt to trick the casual observer into believing something other than what the really data shows. Most misuses are not done intentionally and, without any formal background in statistics, can easily happen. Professional scientists, even statisticians and mathematicians, can be fooled by simplified or improper methodology, despite careful review and proofreading. Scientists have been regularly known to delude themselves with statistics stemming from a poor understanding or education of probability theory and a lack of testing standardization (141–145). Misuse of statistics can be come from many directions including discarding unfavorable data (i.e. cherry picking), overgeneralization with poorly representative sample sets, sample bias, inappropriate estimation of error (e.g. using standard error instead of standard deviation), false causality, and confusing statistical significance with practical significance (152). First and foremost, remember that statistical methods are based on many assumptions which are seldom fully met. An automatic skepticism of results and conclusions is mandatory in science and best practices should always include careful experimental planning, data analysis, drawing logical conclusions, and thorough reporting. Either by intent, or through ignorance or carelessness, statistical fallacies are highly destructive.

1.5 Summary

This chapter has been intended to provide a quick overview of general technologies, methods, and best practices for computational modeling to individuals who may be new to the discipline. Further reading into any of the described methodologies is greatly encouraged. The techniques described in this chapter are largely focused on drug discovery and biochemistry, but can be employed in many fields of study.
A general list, but by no means exhaustive, of popular software that has a comprehensive range of molecular modeling programs is provided below (Table 7). Commercially available software tends to be marginally superior in accuracy and efficiency compared to free or open-source software but differs most in ease of use or access to a larger range of inter-compatible modeling methods. CHARMM and Schrödinger are some of the best in this area, however, Schrödinger software is far more expensive. CHARMM, on the other hand, requires more third party software and significant understanding of the underlying code and UNIX operating systems to access the program’s full potential. Nonetheless, anyone can pull together a comprehensive modeling platform from individual open-sourced programs with enough study and persistence. Further, each tool possesses their own pros and cons, resulting in no single modeling program or methodology being superior in every aspect compared to others. As such, any individual should learn as many different techniques and softwares as possible to ensure success in any modeling endeavor.

Due to the versatile, comprehensive, and integrated nature of its software, as well as available licensing, the majority of this work was based on Schrödinger programs. This introduction centered around the establishing the theoretical framework for the next three chapters, which focus on practical application of the aforementioned molecular modeling techniques in drug discovery projects.
Table 7. Software for computational modeling

<table>
<thead>
<tr>
<th>Program</th>
<th>License</th>
<th>Viewer</th>
<th>Model Builder</th>
<th>MD</th>
<th>MC</th>
<th>QM (Post-HF)</th>
<th>QM/MM</th>
<th>DFT</th>
<th>Semi-Empirical</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMBER*</td>
<td>Freea</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes1</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>CHARMM*</td>
<td>Commercialb</td>
<td>Yes1</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes2</td>
<td>Yes</td>
<td>Yes1</td>
<td>Yes1</td>
<td>Yes</td>
</tr>
<tr>
<td>GROMACS</td>
<td>Open Source</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Limited</td>
<td>No</td>
<td>Limited</td>
</tr>
<tr>
<td>LAMMPS</td>
<td>Open Source</td>
<td>Yes2</td>
<td>Yes2</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Limited2</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>MOE</td>
<td>Commercial</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Limited2</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>NAMD</td>
<td>Open Source</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Limited</td>
<td>Limited</td>
<td>Yes2</td>
</tr>
<tr>
<td>NWChem</td>
<td>Commercial</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Q-Chem</td>
<td>Commercial</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Schrödinger*</td>
<td>Commerciala</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Limited</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>YASARA</td>
<td>Commercialb</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

* Program consists of many individual and possibly separately licensed and developed softwares
a Requires separate modules not included in primary program
b Has interface for required third-party software
c Some software modules have license restrictions
d Limited access or downgraded free version available

1.6 Acknowledgments

The majority of this introduction will be published in Burger’s Medicinal Chemistry, Drug Discovery and Development 8th edition as a chapter on computational chemistry tools. We would like to thank Burger’s Medicinal Chemistry for the opportunity to be involved with the textbook.
2 The Atypical Protein Kinase C Small Molecule Inhibitor ζ-Stat, and Its Effects on Invasion Through Decreases in PKC-ζ Protein Expression

2.1 Introduction

In 2019, it is estimated that approximately 1.7 million new incidences of cancer will be diagnosed and 600,000 cancer deaths will occur (207). Cancer has become the leading cause of death in 21 states, despite decreases in cancer death rates since 1991 (208). Specifically, ovarian cancer is estimated to reach 22,530 diagnoses and cause 13,980 cancer deaths per year (207). This constitutes 5% of cancer deaths among women and is responsible for being the most lethal gynecological cancer diagnosis (209).

The most common ovarian cancer diagnosis is epithelial ovarian carcinomas (EOC), which constitutes 85-90% of diagnosis (209). Of this percent, clear cell ovarian carcinoma (CCOC) represents 5% of incidence and presents unique pathological features and a chemo resistant phenotype (210,211). CCOC is the third most common subtype of ovarian cancer, has a higher risk of reoccurrence, and lower survival rate (210). Furthermore, CCOC is characterized as a non-serous (NS) ovarian cancer and has been found to be more invasive than high grade serous ovarian carcinoma (HGSOC). It has been found that patients with NS tumors have poor prognosis (212). Consequently, CCOC is proposed to be associated with endometriosis and it has been suggested that the pre-cancerous lesions from endometriosis can lead to CCOC (relative risk = 12.4) (213). Endometrial cancer and CCOC have been shown to have an overexpression of
Protein kinase C (PKC) isoforms which play important roles in these cancers development and resistance (214).

Protein kinase C (PKC) is an enzymatic family of proteins that have been found to be a component in cancer progression (215). These proteins phosphorylate the serine and threonine residues of substrates and are generally activated by compounds such as diacylglycerol (DAG), calcium (Ca\(^{2+}\)) and phorbol esters (215). There are three classifications within the PKC family which include the *conventional* PKC-α, βI, βII (splice variant), γ, the *novel* PKC-δ, ε, η, θ, and the *atypical* PKC-ζ, υ,λ (215).

The atypical PKC isoforms, PKC-ι and PKC-ζ, have been suggested to participate in the increased proliferation of ovarian cancer (216). PKC-ι has also been identified as a highly amplified gene in CCOC (210) and is noted for its role in apical-basal polarity loss (216). In addition, due to mutations in the PIK3CA gene and inactivation of Phosphatase and Tensin Homolog (PTEN), the Phosphoinositide 3-Kinase (PI3K)/ Serine Threonine Kinase 1 (AKT)/ Mechanistic Target Of Rapamycin Kinase (mTOR) pathway has also been upregulated in CCOC (211,217–219). The upregulation of this pathway increases the expression of downstream survival targets (e.g. PKC-ζ). PKC-ζ has been shown to be involved in tumorigenesis, tissue invasion, and cancer progression through the modulation of cell migration machinery, such as Ras Homolog Family Member A (RhoA), Rac Family Small GTPase 1 (Rac1), and Epithelial Cell Transforming 2 (Ect2) (220–222).

The ECT2 gene is highly amplified in CCOC and may increase migratory behavior (210). Ect2 is a Rho GTPase specific guanine nucleotide exchange factor (GEF) which activates this family of proteins by the addition of a phosphate group to Guanosine diphosphate (GDP) (223). The overexpression of Ect2 protein promotes increased activation of the Rho GTPases, which in turn can facilitates invasion through cytoskeleton reorganization (224).
Previous studies have indicated that novel aPKC inhibitors ICA-1S and ζ-Stat (Figure 6) decreased the migratory behaviors of colorectal cancer cells and were selective for PKC-ι/λ and PKC-ζ, respectively (222,225). These small molecule inhibitors were also shown to decrease cell viability in colorectal cancer and melanoma (222,225).

![Fig 6](image)

**Figure 6.** Atypical PKC inhibitors. The molecular structures and molecular weights of ICA-1S and ζ-Stat. ICA-1S was synthesized by United Chemistry Resources and ζ-Stat was distributed by the NCI.

Furthermore, computational molecular docking was performed on PKC-ι and a homology model of PKC-ζ (since there is no crystal structure available) with ICA-1S and ζ-Stat (225). In this study, the authors suggested that ICA-1S bound to a potential allosteric pocket (225). However, a more in-depth analysis of ζ-Stat is needed for subsequent studies. The further development of computational modeling is pivotal for drug discovery optimization and helps push these small molecule inhibitors towards a clinical setting. Computational studies can
generate mechanistic understandings of the activity these compounds present, can allow for inhibitor improvement, and can institute further signaling investigations.

It has been suggested that the distal downstream signal cascade of PI3K/aPKC pathway should be targeted due to the genotypic and phenotypic reliance of this pathway in CCOC for survival and invasion. The aims of this study were to further determine the binding mechanisms of ζ-Stat, expand on the tissue range of these compounds by investigating the effects in CCOC cell lines, investigate the therapeutic potential of ζ-Stat in CCOC, and to illustrate the disruption of invasion via the PKC-ζ signaling cascade.

2.2 Materials and Methods

2.2.1 Antibodies and Reagents

The small molecule inhibitors, ζ-Stat and ICA-1S, were obtained from the National Institute of Health (NIH) branch National Cancer Institute (NCI) and United Chem Resources in Birmingham Alabama, respectively. The sources of cell lines, reagents and antibodies were: TOV21G and ES-2 CCOC cell lines (American Type Culture Collection, USA); SHT290 normal endometrial stromal cell line (Kerafast, USA); MCDB 121, Media 199, F12K, penicillin and streptomycin, trypsin, Dulbecco’s phosphate buffered saline (DPBS) and Mito + (Corning, USA); McCoy’s media (HyClone, USA); Opti-MEM I (Gibco, USA); Fetal bovine serum (FBS, Atlanta Biologicals, USA); human insulin (MP Biomedicals, LLC, France); dimethyl sulfoxide (DMSO, Sigma Aldrich, USA); Water-Soluble Tetrazolium (WST-1, Roche, USA); Halt protease and phosphatase inhibitors cocktail and Protein A/G magnetic beads (Thermo Scientific, USA); anti- PKC-ζ (9372s, 1:1000, Cell Signaling, USA); anti- PKC-ι (610178 1:1000, BD, USA); anti-β-actin (A3854, 1:40000, Sigma Aldrich, USA); anti-RhoA (ab54835,1:4000,
Abcam, USA); anti-Ect2 (07-1364, 1:1000, Millipore, USA); Donkey anti-rabbit IgG Alexa-488 (A21206, 1:500, Invitrogen, USA); Goat anti-rabbit (170-6515, 1:2000, Bio-Rad Laboratories, USA); Goat anti-mouse (170-6516, 1:2000, Bio-Rad Laboratories, USA); Activated Rac1 pulldown kit (BK035, Cytoskeleton, USA); 96-well transwell insert and basement membrane extract (BME; both Corning Inc., Corning, NY, USA); RNA bee (Amsbio, United Kingdom); Qiagen RT Kit (205113, Qiagen, Germany); RhoA PCR primers (HP100025, Sino Biological, USA); Glyceraldehyde 3-phosphate dehydrogenase (GAPDH) PCR primers (Eurofins, USA)

2.2.2 Analysis of Somatic Gene Mutations For Ovarian Cancer, CCOC and CCOC Cell Lines

The selection of CCOC cell lines employed the COSMIC database (226). Initially, the search was focused on all ovarian subtypes in the database. The search was then re-focused on clear cell carcinomas. Furthermore, the COSMIC cell line project was utilized for the analysis of mutations in TOV21G and ES-2 cells (226).

2.2.3 Computational Analysis of aPKCs

2.2.3.1 Protein Preparation. Protein model systems for PKC-ι and PKC-ζ were prepared using the Schrodinger software suite (25). Protein structure coordinates were downloaded from the Protein Data Bank (PDB) (227, 228). The PKC-ι model was generated from the PDB 3A8W entry co-crystallized with adenosine triphosphate (ATP) (229). The apo structure, PDB 38AX, was used to cross reference conformational states from Molecular Dynamics (MD) simulations. No crystal structure currently exists of PKC-ζ, necessitating the need for homology modeling to attempt to produce a potentially viable docking model for PKC-ζ. Two PKC-ζ models were built, one utilizing SWISS-MODEL (19, 20, 230) and another employing the Prime homology program.
(16,66,67) using the human PKC-ζ sequence (UniProt Q05513) and the crystallized PKC-ι structures as templates. PDB systems were prepared with the Protein Preparation Wizard (PrepWizard) in Maestro (231,232). Cofactors used in crystallization (such as sulfate or phosphate ions), ligands, and additional protein dimers were deleted. Bond orders were then assigned, including disulfide bridges, and original hydrogens were deleted and later replaced to reduce bad contacts and other crystal artifacts before protonation and hydrogen bond optimization. All waters were retained for assisting in the determination of side chain protonation states and initial hydrogen bond optimization. Missing side chains were added and optimized using Prime. Hydrogen atoms were then added to the protein, remaining cofactors, and to any added structural waters. The program PROPKA (50) was used for the prediction of protein ionization states at 7.4 pH and ProtAssign was used for hydrogen bond optimization. After automatic hydrogen assignment, visual inspection was used to flip residues and change protonation states at protein-protein interfaces if and when appropriate.

2.2.3.2 Molecular Dynamics. MD simulations were performed with the Desmond MD program (233–236). A cubic simulation box was created extending at least 10Å from the protein with imposed periodic boundary conditions. TIP3P waters (237) were added to solvate the simulation box and was then electrically neutralized by introducing sodium ions. The OPLS-3 all-atom force field (34) was then applied to all atoms. The SHAKE algorithm (238) was used to constrain all bonds in the system and the REference System Propagator Algorithm (RESPA) (239) with an integration time step of 2 fs was employed. The Particle Mesh Ewald (PME) algorithm was used to calculate long-range electrostatics with a real-space cutoff of 13 Å. Van der Waals interactions were cutoff at 16 Å. The systems were simulated in an NPT ensemble using the Nose–Hoover temperature coupling scheme (240) at a temperature of 310 K and a
constant pressure of 1 atm using the Martyna-Tuckerman-Tobias-Klein (MTTK) barostat (241,242).

All systems were energy minimized with a truncated newtonian conjugate gradient (TNCG) method (243) followed by multiple restrained minimizations to randomize systems before equilibration and final simulation. Heavy atoms of the protein were held fixed during heating for an initial 12 ps NVT ensemble simulation at 10 K with the Berendsen thermostat (244). This was followed by simulations at 1 atm in the NPT ensemble for 12 ps at 10 K and 24 ps at 310 K. Unrestrained equilibration MD was then performed for 24 ps at 310 K and 1 atm. Finally, unconstrained production MD was performed on PKC-ι and PKC-ζ systems for 250 ns. Energies were recorded every 2 ps and trajectory frames were recorded every 5 ps.

Final system equilibration was determined by the observation of asymptotic behavior of the potential energy, Root Mean Square Deviation (RMSD), and Radius of Gyration (Rg) profiles and visual inspection of trajectories guided by Root Mean Square Fluctuation (RMSF) profiles (Figures 8-11).

**2.2.3.3 Consensus Docking.** After equilibration was determined, a hierarchical average linkage clustering method based on RMSD was utilized to determine an average representative structure for the equilibrated PKC-ι system. The program PROPKA was then implemented again on the equilibrated structure to test consistency of side chain protontion states at 7.4 pH. The representative structure was then used for consensus docking incorporating five diverse and complimentary docking methods described below. By applying these varied energy scoring methods, the weaknesses of each method can be identified for a particular model and error statistically minimized, yielding a more accurate summary of ligand binding dispositions and affinities.
As a check for the placement of the grids used in the docking studies and for further analysis of the binding cavities for the ATP binding site and the potential allosteric site, Schrödinger's SiteMap program (245–247) was employed. SiteMap searches the protein structure for likely binding sites and highlights regions within the binding site suitable for occupancy by hydrophobic groups, hydrogen-bond donors, acceptors, or metal-binding functionality of the ligand.

The ligands ICA-1S, ζ-Stat, and ATP were prepared using the program LigPrep (248) and the OPLS-3 all-atom force field was applied to all ligand atoms.

### 2.2.3.4 Rigid Receptor Docking (RRD).

Rigid docking simulations were performed by Glide (114,121,125). Glide uses a GlideScore fitness function based on Chemscore (249,250) for estimating binding affinity, but includes a steric-clash term, adds buried polar terms to penalize electrostatic mismatches, and modifies other secondary terms.Docking simulations used both the standard precision (SP) and extra precision (XP) methods. XP mode is a refinement algorithm enforced only on good ligand poses. Sampling is based on an anchor and refined growth strategy and the scoring function includes a more complete treatment of some of the SP energetic terms, such as the solvation and hydrophobic terms. Docking grids were defined by a rectangular ligand atom inclusion outer box of 22Å and ligand centroid constraint inner box of 10Å in the x, y, and z directions originating from the binding cavity centroid defined by SiteMap for the proposed allosteric site and by the original co-crystallized ATP ligand centroid for the ATP binding site.

### 2.2.3.5 Induced Fit Docking (IFD).

The IFD methodology (16,122–125) incorporates both the docking program Glide to account for ligand flexibility and the Refinement module in the Prime program to account for receptor flexibility. The Schrödinger IFD protocol attempts to
model induced-fit effects from alterations in binding site conformation due to ligand binding in order to increase accuracy of binding affinity estimates and prediction of possible binding modes.

The position of the cubic docking grid for the ATP binding site was centered on the original co-crystallized ligand centroid and from the binding cavity centroid defined by SiteMap for the proposed allosteric site with a box size of 29 Å for both. A constrained minimization of the receptor was performed with an RMSD cutoff of 0.18 Å. An initial softened potential Glide docking of the ligand set was then implemented with the standard precision (SP) mode and a van der Waals scaling factor of 0.5 was applied to the non-polar atoms of the receptor and ligands. The resulting top 20 poses of the ligands were used to sample protein plasticity by conformational searches and minimizations of binding pocket residues within 6 Å of any ligand pose for all complexes obtained. The new receptor conformations were then redocked using complexes within 30 kcal/mol from the best scoring structure. Glide docking parameters for this step were reset to the default hard potential function with a van der Waals scaling of 1.0 and SP mode.

The estimated binding affinity of each complex was reported in the GlideScore and used to compare differences between each ligand while the Emodel score is used to inter-compare poses of the ligands. Emodel places more significance on weighting force field components (electrostatic and van der Waals energies), making it better for comparing conformers as opposed to comparing chemically-distinct species.

2.2.3.6 Quantum Polarized Ligand Docking (QPLD). To account for ligand polarization upon binding, Quantum Mechanics/Molecular Mechanics (QM/MM) docking was performed by the Schrödinger QM-Polarized Ligand Docking Protocol (QPLD) (126,127). The
protocol first employs RRD using Glide in SP mode. In this step, the top five poses of each ligand in the initial RRD were used. Potential ligand polarization induced by the protein were then calculated with Qsite \((111,251,252)\) at the B3LYP/6-31G* level. The ligand force fields were then reconstructed with QM/MM modified charges, redocked, and five poses of each ligand were saved for evaluation.

2.2.3.7 Molecular Mechanics and Generalized Born Surface Area (MM/GBSA). The MM/GBSA method combines molecular mechanics energy terms and implicit solvation models to calculate the binding-free energy based on docking complexes. The protocol, implemented by the Prime MM-GBSA module, calculates optimized free energies for the free protein and free ligand and references them with the original bound complex energy \((253)\). Polar contributions are calculated using the Generalized Born (GB) model \((254)\), an implicit solvent model is based on a variable dielectric surface Generalized Born (VD-SGB) approach, where the variable dielectric value for each residue was fit to a large number of side-chain and loop predictions while the non-polar energy is estimated using the solvent accessible surface area (SASA) \((255)\). The simulation was performed based on receptor–ligand complex structures obtained from induced fit docking. The obtained ligand poses were minimized using the local optimization feature in Prime, whereas the energies of complex were calculated with the OPLS-3 force field and Generalized-Born/Surface Area continuum solvent model \((256)\). During the simulation process, the ligand strain energy is also considered. A known issue with MM/GBSA is that scores do not accurately reproduce absolute physical binding affinities but display great efficacy at ranking compounds in a relative manner \((257–260)\). We developed a correlation function using a single-layer logistic regression to rescale MM/GBSA scores based on the other docking score.
algorithms. This retains the ranking accuracy of MM/GBSA and allows us to proportion the results in a minimally biased and physically relevant manner.

2.2.3.8 Virtual Target Screening (VTS). VTS is a system designed to virtually screen a molecule of interest to a large library of protein structures. The current protein library consists of 1,451 structures with a concentration of kinases. The system is calibrated with a set of small drug-like molecules are docked against each structure in the protein library to produce benchmark statistics. VTS was employed as a theoretical assay of potential kinase activity and gauge of potential biological promiscuity. The calibration procedure allows the analysis to accurately predict inhibitor–kinase binding affinities when $K_d < 10 \, \mu M$ (defining a hit) and $K_d \geq 10 \, \mu M$ are both considered (72% accuracy in the best case) (137). Therefore, the VTS system is able to robustly discriminate protein binders from nonbinders and give some inclination as to potential binding promiscuity of the molecule of interest with respect to the protein group tested.

2.2.4 Cell Culture

The CCOC cells lines TOV21G and ES-2 were cultured in MCDB 131: Media 199 (1:1 ratio) and McCoy’s medium, respectively, supplemented with 10% FBS, 100 units/mL of penicillin and 100μg/mL of streptomycin. The immortalized normal human endometrial stromal cell line, SHT290, was maintained in F12K: Media 199 (1:1 ratio) and supplemented with 5% FBS, 0.1% Mito +, 2μg/mL of human insulin, 100 units/mL of penicillin and 100μg/mL of streptomycin. All cell cultures represented were passaged less than 10 times. Cell cultures were maintained in an incubator at 37 °C and 5% CO$_2$ atmosphere.
2.2.5 Atypical PKC Expression During Rapid Growth and Cell Cycle Arrested

SHT290, TOV21G and ES-2 cells were seeded into 100mm plates and grown to 50% confluency. Rapidly growing cells were harvested at 50% confluence and the cell cycle arrested cells were serum starved for an additional 48 hrs in reduced serum medium Opti-MEM I (N=3).

2.2.6 Preliminary Screening of TOV21G and ES-2 with ICA-1S and ζ-Stat

TOV21G and ES-2 cells were seeded into 100mm plates and grown to 50% confluency. Control cells were treated with equal amounts of DMSO (vehicle) and treatment cells were treated with 3µM concentrations of ICA-1S and ζ-Stat for 24, 48 and 72 hrs (N=3). Cells were collected at 24, 48 and 72 hrs and run on a Western blot.

2.2.7 Cell Viability Assay

Cells were seeded in to 96-well plate at 800 cells per well with 200µL of media. Cells were treated with different concentrations of DMSO (vehicle to match treatment, N=12) and ζ-Stat (1µM, 3µM, 5µM and 10µM, N=3). After 72 hrs of treatment, the cell viability was analyzed using WST-1 at wavelengths 450 and 630 nm. The plates were read on a BioTek SynergyHT microplate reader. Standard curves for each cell line was generated based on the number of cells added and the absorbance recorded.

2.2.8 Cell Lysate Collection

Media was extracted from the vessel and 250µL of lysis buffer [Pierce® Immuno Precipitation Lysis Buffer, 25 mM Tris-HCl pH 7.4, 150 mM NaCl, 1 mM EDTA, 1% NP-40 and 5% glycerol] with protease and phosphatase inhibitors was added to the plates. Cells were scraped and collected from the vessel (on ice) and the suspension was sonicated for 3 x 5 s cycles on ice. The samples were centrifuged at 4°C at 12,000 x g for 15 ms. The supernatant (cell
lysate) was removed from the cellular membrane pellet and placed in a secondary micro centrifuge tube. Protein content was measured per Bradford Assay Reagent on a BioTek SynergyHT microplate reader at 595nm.

2.2.9 Western Blot Analysis

Cell lysates containing equal amounts of protein (20-40µg) were loaded in each lane and run on sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) and transblotted to a 0.45µm nitrocellulose membrane. Cell lysates were probed with the primary antibodies against PKC-ζ, PKC-ι, RhoA and β-actin (for loading control) and re-probed with secondary antibodies for development. Immunoreacted bands were visualized by enhanced chemiluminescence per the manufacturer’s instructions [Thermo Scientific™ SuperSignal™ West Pico PLUS Chemiluminescent Substrate]. Band densitometry was performed with ImageJ FIJI (261) software and normalized densities were derived by the ratio of the protein of interest over the control (β-actin).

2.2.10 Endpoint PCR

The mRNA from TOV21G and ES-2 cells treated with vehicle control and 3μM of ζ-Stat was isolated via the manufacture’s protocol for RNAbee. The mRNA was then quantified using the BioTek Synergy HT Take5 plate and cDNA was synthesized using the Qiagen reverse transcription (RT) kit per the manufacturer’s protocol. Endpoint polymerase chain reaction (PCR) was performed with primers for RhoA and the house keeping gene GAPDH (forward 5’CCA-CCC-ATG-GCA-AAT-TCC-ATG-GCA-3’ and reverse 5’TCT-AGA-CGG-CAG-GTC-AGG-TCC-ACC-3’). PCR products were analyzed on an agarose gel (N=3).
2.2.11 Rac1 Activation Assay

Cells were cultured in 150mm plates and lysed as previously described. The cells were treated with DMSO (control) and 10μM of ζ-Stat for 72 hrs and harvested (N=4). Glutathione S-transferase (GST) was tagged to the protein binding domain (PBD) of p21 activated kinase (PAK). A positive control and negative control were performed to determine assay efficiency. Briefly, 500μg of protein were balanced in 200μL of cell lysis for each sample. The positive control received 200μM of non-hydrolyzable guanosine 5'-(gamma-thio) triphosphate (GTPγS) and the negative control received 200μM of guanosine diphosphate (GDP). These samples were incubated at room temperature (RT) for 15 ms. All samples (positive, negative, DMSO control and treatment) were incubated with GST-tagged PAK-PBD agarose beads for 1 hr 4°C. These samples were pelleted at 5,000 x g (at 4°C) and washed with Wash Buffer. The pelleted beads were re-suspended with 20μL of 2X Laemmli sample buffer and boiled at 95°C for 2 ms.

2.2.12 Preparation of cytoplasmic and nuclear extracts

TOV21G and ES-2 cells were seeded in 100mm tissue culture plates (1.5x10⁵). Cells were treated for 72 hours with 10μM ζ-Stat (DMSO control) and harvested with trypsin. The instructions provided by the manufacturer were followed to fractionate the cytoplasmic and nuclear portions. The extracts were analyzed via immunoblots and translocation of Ect2 was investigated.

2.2.13 Fluorescent Microscopy

TOV21G cells were seeded into 4 chambered slides at a 500 cells per well concentration and after 24 hrs, were treated with a vehicle control (DMSO) and 10μM of ζ-Stat every 24 hrs
for 72 hrs. Cells were then fixed with 4% paraformaldehyde for 15 ms and immunostained with Ect2 antibody at 4°C overnight with light agitation. The slides were incubated with Alexa 488 rabbit secondary antibody for 1 hr at room temperature RT. Subsequently, the slides were stained with Phalloidin conjugated to Texas red dye for 30 ms at RT, mounted with solution containing 4’,6-diamidino-2-phenylindole (DAPI) and imaged on an Olympus BX53 Digital Upright Fluorescent Microscope.

2.2.14 Invasion Assay

For the evaluation of invasion, cells were serum starved for 48 hrs, followed by detachment and plating into the upper chamber of a 96-well (8 µm) transwell permeable support, coated with 0.1X BME. Serum (10%) containing media was loaded into the lower chamber as a chemoattractant. Subsequently, TOV21G cells at the upper chamber were treated with 10µM of ζ-stat for 24 hr (N=4). Two experimental treatment groups for the cells were performed: Control (DMSO vehicle) and treatment. The invasive cells that passed into the lower chamber were then fixed with 4% paraformaldehyde, stained with 2% crystal violet in 2% ethanol, washed with distilled water and photographs were captured after drying using a light microscope Motic AE31E. For migration, a similar protocol was followed except without coating the transwell insert with BME. The assay was quantified with ImageJ FIJI software.

2.2.15 ζ-Stat in-vivo

The following experiments outline the investigations of ζ-Stat in TOV21G clear cell carcinoma ovarian xenografts. We have an Institutional Animal Care and Use Committee (IACUC) approved by Adrienne Booker for the discussed studies. The study involved 12 athymic female nude mice weighing between 20-25g and >10 weeks of age. The 12 mice were
divided into two groups after TOV21G cells were implanted (1 x 10^6 cells/per mouse flank in 0.2mL of media). The first group was the vehicle control group (N=6), which received 100 µL of 1x DPBS. The second group (N=6) was injected with 100 µL of 20mg/kg of ζ-Stat dissolved in 1x DPBS. The tumor volume was calculated using the formula: length x width x width x ½.

Three days after the implantation of the cells, tumors were treated as of day 0. The treatments were administered every other day subcutaneously intra-tumor and around the tumor site for 35 days.

At the end point of the experiment, tumors and heart serum were harvested. Tumors were imaged and measured, and blood serum was analyzed for enzymatic levels of glucose (GLU), blood urea nitrogen (BUN), alanine aminotransferase (ALT), aspartate aminotransferase (AST) and alkaline phosphatase (ALKP) at the Moffitt Research Facility. Briefly, blood chemistry analysis was performed by initially collecting whole blood in a serum separator tube, which then sat for 20 ms at RT before centrifugation. Once the blood was centrifuged the serum was separated and placed in a specialized sample cup made for the IDEXX CatalystDx. The cup containing the serum and the desired chemistry slides were then placed into the CatalystDx for analysis.

2.2.16 Statistical Analysis

R studio software was used for statistical analyses. A one-way ANOVA with a Tukey’s multiple comparisons test was performed for Western Blot analyses and cell viability. A two-tailed unpaired student T-test was utilized for the statistical significance of the particle counts for cell migration and invasion, day to day tumor volume, mouse body weight and individual enzyme levels. The Pearson’s Correlation Coefficient (PCC) was utilized for co-localization and was analyzed using ImageJ FIJI software, using the Coloc2 plugin.
2.3 Results

2.3.1 PIK3CA and ARID1A Are in the Top Mutated Genes in All Ovarian Tissue Types and in CCOC

To understand the genetic landscape of ovarian cancer, we utilized the Catalogue of Somatic Mutations in Cancer bioinformatics database (COSMIC) (226). The results demonstrated that the top mutations in ovarian cancer overall are TP53 (p53), FOXL2 (Forkhead box protein L2), KRAS (Kirsten Ras oncogene homolog), PIK3CA (Phosphatidylinositol 4,5-bisphosphate 3-kinase catalytic subunit alpha), ARID1A (AT-rich interaction domain 1A) and BRAF (B-Raf proto-oncogene) (Table 8). The search was then refocused to only contain CCOC samples and the top two mutated genes found were PIK3CA (33%) and ARID1A (50%) (Table 9). These results suggest that one of the most common gene mutations in ovarian cancer and CCOC is PIK3CA, which is approximately 10% of mutated samples in all the ovarian tissue in the database. Due to this, the downstream survival targets PKC-ζ and PKC-ι are likely to be overexpressed, amending the need for their explicit targeting.

In addition, the PIK3CA mutation was used to select two cell lines that would be representative of this mutation in CCOC. Two commonly utilized cell lines, TOV21G and ES-2 were selected based on their genetic profile. While both cell lines possess a PIK3CA mutation, TOV21G has an ARID1A mutation and ES-2 has a TP53 mutation (Table 9).
**Table 8. Six most common gene mutations in all ovarian cancers.**

<table>
<thead>
<tr>
<th>Gene</th>
<th>Protein Product</th>
<th>Protein Function</th>
<th>Chromosomal Location (human)</th>
<th>Percent Mutation of Samples Tested</th>
<th>Highest Percent Mutation</th>
<th>Mutation Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP53</td>
<td>p53</td>
<td>Tumor suppressor, regulates cell cycle</td>
<td>17p13.1</td>
<td>46</td>
<td>55.89, 20.23</td>
<td>Substitution missense, other</td>
</tr>
<tr>
<td>FOXL2</td>
<td>Forkhead box protein L2</td>
<td>Transcription factor</td>
<td>3q23</td>
<td>20</td>
<td>100</td>
<td>Substitution missense</td>
</tr>
<tr>
<td>KRAS</td>
<td>Kirsten Ras oncogene homolog, (KRAS proto-oncogene, GTPase)</td>
<td>Regulation of cell division</td>
<td>12p12.1</td>
<td>12</td>
<td>100</td>
<td>Substitution missense</td>
</tr>
<tr>
<td>PIK3CA</td>
<td>Phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha</td>
<td>Phosphorylates certain signaling molecules</td>
<td>3q26.3</td>
<td>10</td>
<td>96.88</td>
<td>Substitution missense</td>
</tr>
<tr>
<td>ARID1A</td>
<td>AT-rich interaction domain 1A</td>
<td>Regulate transcription by altering chromatin structure</td>
<td>1p35.3</td>
<td>9</td>
<td>40, 38.26, 23.48</td>
<td>Substitution missense, deletion frame shift, insertion frame shift</td>
</tr>
<tr>
<td>BRAF</td>
<td>B-Raf proto-oncogene, serine/threonine kinase</td>
<td>This protein plays a role in regulating the MAP kinase/ERKs signaling pathway, which affects cell division, differentiation, and secretion.</td>
<td>7q34</td>
<td>7</td>
<td>97.63</td>
<td>Substitution missense</td>
</tr>
</tbody>
</table>

The table describes the six most common gene mutations in all ovarian cancers according to the Catalogue of Somatic Mutations in Cancer database (COSMIC). The gene name, protein product, general function, chromosomal location, percent of all samples with mutation, highest type of mutation and most common mutation type are listed.
### Table 9. Somatic mutations in CCOC and CCOC cell lines.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Percent Mutation of Samples Tested</th>
<th>TOV21G Mutations</th>
<th>ES-2 Mutations</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARID1A</td>
<td>50</td>
<td>ARID1A&lt;sub&gt;Y551fs&lt;/sub&gt;,&lt;sub&gt;Q758fs&lt;/sub&gt;,&lt;sub&gt;Q758fs&lt;/sub&gt;</td>
<td>TP53&lt;sub&gt;S241F&lt;/sub&gt;</td>
</tr>
<tr>
<td>PIK3CA</td>
<td>33</td>
<td>PIK3CA&lt;sub&gt;H1047Y&lt;/sub&gt;</td>
<td>PIK3CA</td>
</tr>
<tr>
<td>TERT</td>
<td>17</td>
<td>KRAS&lt;sub&gt;G13C&lt;/sub&gt;</td>
<td></td>
</tr>
<tr>
<td>TP53</td>
<td>11</td>
<td>PTEN&lt;sub&gt;K267fs&lt;/sub&gt;,&lt;sub&gt;G143fs&lt;/sub&gt;</td>
<td></td>
</tr>
<tr>
<td>KRAS</td>
<td>8</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The table describes the five most common gene mutations in CCOC according to the Catalogue of Somatic Mutations in Cancer database (COSMIC). The gene name and percent of samples with mutation are listed. The gene mutations for TOV21G CCOC cells and ES-2 CCOC cells are listed with specific mutation type.

### 2.3.2 Select Inhibitor Effects on PKC-ζ and PKC-ι Protein Expression

To determine which inhibitors affect PKC-ζ and PKC-ι protein expression in CCOC, Western Blots were employed with ICA-1S and ζ-Stat (Figure 6) as potential inhibitors. Initially, the expression of these aPKCs were investigated in rapidly growing cells (50%) and cell cycle arrested (serum free, SF) (Figure 7A). The density of each band was quantified using analytical software. The results showed that ICA-1S did not affect PKC-ζ or PKC-ι protein expression (Figure 7B); however, ζ-Stat substantially decreased the expression of PKC-ζ in TOV21G cells (<i>p</i> value 0.00225, <i>F</i> = 9.5709, <i>t</i> = -4.413) and ES-2 cells (not significant) but not PKC-ι protein expression (Figure 7C). These results suggest that ζ-Stat is selective to PKC-ζ decreased expression and could be used to interrupt the PKC-ζ pathways.
Figure 7. PKC-ζ and PKC-ι protein expression in rapidly growing and serum starved cells and the effects of ICA-1S and ζ-Stat on PKC-ζ and PKC-ι protein expression in SHT290, TOV21G and ES-2 CCOC cell lines. (A) SHT290, TOV21G and ES-2 cell lines were harvested at 50% and 48 hrs after serum starvation (SF) (N=3). The membranes were probed with anti-PKC-ζ and PKC-ι to investigate protein expression. (B) TOV21G and ES-2 cell lines were treated with 3μM ζ-Stat for 24, 48 and 72 hrs. An untreated control and vehicle control (DMSO) are also illustrated. The immunoblots were probed with PKC-ζ, PKC-ι and β-actin (loading control) (N=3). (C) The raw data’s densitometry was quantified using ImageJ software and analyzed with a one-way ANOVA. Standard deviation is represented. (** p < 0.01)

2.3.3 In-silico Results and Model Validation

2.3.3.1 Model Viability. RMSD and R<sub>g</sub> plots of PKC-ι (Figure 8) displayed asymptotic behavior beyond 7 ns. RMSD appeared to equilibrate near 2.3 Å, similar to the reported crystallographic resolution of 2 Å (229). Any additional fluctuation is likely from the disordered tail regions and missing residues 446 – 454 as evidenced from the PKC-ι RMSF plot (Figure 9).
From these analyses, the MD model is considered viable due to the ability of the model to accurately maintain the lowest energy structure under significant perturbation.

**Figure 8.** RMSD (Bottom) of backbone atoms and Radius of Gyration ($R_g$, Top) of all atoms graphs for PKCι Molecular Dynamics simulation. Equilibration appears to occur after 7 ns with an average RMSD of 2.3 Å.
To date, no crystallographic structure exists of PKC-ζ. As such, a homology model was attempted using the Schrödinger program Prime (16,66,67) with the human PKC-ζ sequence (UniProt Q05513) and the crystallized PKC-ι structures as templates. PKC-ι was chosen as the primary template due to its highest sequence identity (49.0%) and homology (53.8%) given from a BLAST search (205) coupled with highest overall structural resolution. A pre-generated
SWISS-MODEL (19,20,230) homology model was used as a control for structural reference. The Prime model (Figure 12C) was identical to the SWISS-MODEL variant, so only the Prime model was used for further analysis. RMSD and $R_g$ plots of the PKC-ζ homology model (Figure 10) do show asymptotic behavior after 1.2 ns, however the RMSD equilibration is averaged to 4.4 Å and is clearly outside any acceptable resolution. The $R_g$ plot also shows a significant expansion indicative of unfolding with internal water infiltration. RMSF plots (Figure 11) also show substantial backbone movement beyond 2 Å at regions of PKC-ζ residues differing from PKC-ι with notable disruption of predicted secondary structure. These factors conclude that the homology model does not represent a physical low energy structure and therefore cannot be used for further modeling. Despite this, the ATP binding region is largely stable with RMSF well within reasonable values (Figure 12D). This allows inclusion of the PKC-ζ homology model ATP binding site with docking studies to compare binding modes between PKC-ι and PKC-ζ ATP binding sites (Figure 12C – yellow region).
Figure 10. RMSD (Bottom) of backbone atoms and Radius of Gyration ($R_g$, Top) of all atoms graphs for PKCζ Molecular Dynamics simulation. Some asymptotic behavior does appear to occur, however, an average RMSD of 4.5 Å and increasing $R_g$ implies unfolding and significant deviation from any structural energetic minimum.
Figure 11. RMSF (Root Mean Square Fluctuation) plot for PKC-ζ Molecular Dynamics simulation. Graph shows RMSD per residue of α-carbon over length of simulation. The ATP region, approximately residues 300 – 450, show displacements under 2 Å. Most homologous residues show RMSD values well above 2 Å.
Figure 12. Computational modeling results. Binding site characterization for PKC-ι ATP site with co-crystallized ATP shown for reference (A) and the PKC-ι potential allosteric site with ζ-Stat docking result shown for reference (B). SiteMap analysis for each site is also shown (A and B). Yellow volume denotes hydrophobic regions, red volumes demarcate hydrogen bond donor sites, and blue volumes define hydrogen bond acceptor sites. The final homology model structure of PKC-ζ (C) generated is displayed with cartoon ribbons and colored by sequence identity to the PKC-ι reference structure. Blue are identical residues between both structures with cyan highlighting dissimilar amino acids. Red indicates sequence spans with no structural template. The ATP binding region is also delineated by a yellow domain and a magenta domain localizes the suspected allosteric site, both defined by SiteMap. Post-MD RMSD analysis (D) is illustrated with cartoon ribbons colored by RMSD per residue over the length of the simulation. Continuous color shift is used to show residue displacement from original structure with blue representing a minimum of 0 Å RMSD, white representing mid values of 4 Å RMSD, and red representing a maximum of 8 Å RMSD. Consensus docking scores (E-G) are measured in kcal/mol and represent estimated free energies of binding (ΔG°). Green highlighting features the best scoring molecule for a particular site and yellow highlight focuses the averaged consensus scoring for each compound. Theoretical K_D values are calculated from free energies of binding (ΔG°). VTS results are also listed (H).
The Schrödinger program SiteMap was used for optimizing placement of grids used in the docking studies and to search for any other potential binding sites. A possible allosteric site for PKC-ι was identified by SiteMap near residues 397 – 400 in the activation segment of the PKC-ι c-lobe consisting of a pocket made by α-helices E and F and the interhelical loop between α-helices H and I (Figure 12C – magenta region). SiteMap also scores regions based on potential hydrogen bonding, hydrophobicity, and pocket volume. Scores of 0.8 or greater are considered the cutoff for distinguishing between drug-binding and non-drug-binding sites. The PKC-ι ATP site was scored at 1.004 and the potential allosteric site was scored at 0.779. The potential allosteric site score is notably within SiteMap calibration error and was still included in docking studies due to proximity of the pocket with the PKC-ι activation segment and for the possibility of induced fit effects opening the site.

2.3.3.2 Docking Results. Minimal direct binding data exists for PKC-ι and PKC-ζ so a consensus docking approach was utilized to gauge the optimal docking algorithm for the ATP and potential allosteric site. The approach detailed utilizes five different computational methods of discerning theoretical binding affinities: two unique scoring functions (SP and XP) for ridged docking methods, IFD to account for potential induced fit effects, QPLD can resolve polarization effects through QM/MM techniques, and MM/GBSA is superior in clarifying penalties for solvent interactions. Employing these functions when little empirical evidence exists to correlate results helps identify weaknesses of each technique for a particular model. Error can also be statistically minimized, yielding a more accurate summary of ligand binding dispositions and affinities.
Staurosporine was used as a docking control for the ATP site since binding data exists for both PKC-ι (261 nM Kᵢ (262), values converted from IC₅₀s) and PKC-ζ (131 nM Kᵢ (263), values converted from IC₅₀s). ATP is not used as a docking control due to poor model forcefields and lack of direct binding data.

Docking results are summarized in Figure 12E, F, and G. Docking scores for staurosporine controls are well within reasonable agreement with literature values. Docking scores and poses for each molecule are nearly identical for the ATP site of PKC-ι and PKC-ζ and both prefer staurosporine by a significant margin (Figure 12E and F, green highlight). XP scoring consistently yielded scores in closest agreement to literature values and highest Pearson correlation to overall averages. IFD and QPLD have poorer correlation and control accuracy, suggesting a less pronounced influence of charge factors and induced fit effects since including polarization and site flexibility does not increase docking accuracy. MM/GBSA scores exhibit similarly reduced correlation and high variance, entailing that solvent effects are also not likely a major factor for binding; this is understandable given the pocket depth. These analyses signify that the hydrophobic centers of the site (Figure 12A) are the dominant factors in ligand binding with the ATP site for both models. As opposed to the pan-kinase inhibitor, staurosporine, ICA-1S and ζ-Stat display negligible binding with the ATP site.

The potential allosteric site was also studied (Figure 12G), but only for PKC-ι since the corresponding PKC-ζ allosteric site model could not be validated. As such, no conclusions should be drawn concerning possible interactions of these compounds with any potential allosteric sites on PKC-ζ (only the ATP site achieved an apparent suitable stability for docking studies). Of the three molecules, the potential allosteric site appears to prefer only ICA-1S with a theoretical Kᵦ of 1.4 μM. A recent study by Ratnayake et al. (225) measured myelin basic protein
(MBP) phosphorylation by PKC-ι and PKC-ζ in the presence of ICA-1S, ICA-1T (the phosphorylated version of ICA-1S), and ζ-Stat. Docking figures match expected activity of PKC-ι for ζ-Stat having negligible inhibition. ICA-1S activity as a PKC-ι inhibitor gives some support to the existence of the potential allosteric site, since modeling suggests that ICA-1S does not interact significantly with the ATP site, but does display binding at concentrations similar to experimental values for inducing inhibition. Unfortunately, the data for PKC-ζ is less clear. All the modeling can show is if the two compounds in question can effectively bind to the ATP site of PKC-ζ. This may indirectly imply an allosteric site exists for PKC-ζ if inhibition is experimentally observed and the compound in question does not appear to have favorable docking to the ATP site. For ICA-1S, there is negligible affinity for the ATP site and experiment reflects a lack of inhibition. Modeling also suggests an allosteric mechanism may be present for ζ-Stat as binding is also negligible for the ATP site. Experimental inhibition should be observed for any compounds that compete for the ATP site since there is no significant difference between the ATP sites of PKC-ι and PKC-ζ.

2.3.3.3 Virtual Target Screening Results. VTS uses a large curated protein structure library to which molecules of interest are docked. Statistical calibrations and baselines are applied to average and relate docking scores with each individual and class of proteins. A kinase-enriched library (1,451 proteins, 464 transferases, and 65 unique kinases) was assigned for docking with the three compounds. A hit on a protein is classified as the potential ($p$ value < 0.05) of the molecule of interest to bind to the specified protein with a theoretical $K_D$ of 10 μM or less. This analysis can infer the specificity of a molecule for a particular class of proteins. The VTS results for each compound are listed in Figure 12H.
The staurosporine control gave an expected baseline commensurate of a pan-kinase inhibitor. It displayed low to moderate total protein activity with moderately high interaction with general transferases and hit a majority of kinases, alluding a clear preference for kinases. ICA-1S demonstrated a low total protein activity with a slight but pronounced increase in affinity for transferases and kinases. This suggests a possibility of seeing some expected broad kinase interference for ICA-1S. ζ-Stat, however, portrays significant specificity in VTS. It has similar low hit percentages for all protein classes, implying little to no expected kinase activity.

2.3.4 Inhibition of Cell Viability

The effects of ζ-Stat on CCOC cellular viability was investigated via WST-1 methodologies. The results revealed that 10μM ζ-Stat did not significantly effect SHT290 normal endometrial stromal cells, but did significantly decreased the viability by 37% in TOV21G cells ($p$ value 0.0436, $F = 4.2461$, $t = -3.058$) and by 57% in ES-2 cells ($p$ value 0.00363, $F = 7.2918$, $t = -4.220$) (Figure 13A-C). These results suggest that ζ-Stat decreases the viability of CCOC but has negligible effects on normal endometrial stromal cells.
2.3.5 Analysis of the PKC-ζ/Ect2/Rac1/RhoA Pathway

To determine the downstream effects of ζ-Stat on invasion, immunofluorescence, Western Blots and semi-quantitative endpoint PCR techniques were utilized. Vehicle control and ζ-Stat treated TOV21G cells were probed with anti-Ect2 and imaged. The results showed that Ect2 was present in the filamentous extensions in control cell. After treatment, the polarity of the filamentous extensions decreased and Ect2 was found to be more abundant around the nucleus (Figure 14A-B). Western results demonstrated that 3μM of ζ-Stat decreased RhoA protein.
expression as well as mRNA expression (Figure 14C-D). These results suggest that the decrease in PKC-ζ protein expression reduces the expression of RhoA at the genomic level.

Furthermore, Ect2 localization was observed by immunofluorescence with and without treatment. The PCC showed that the control had a lower amount of Ect2 nuclear localization (0.57) in comparison to the treated (0.72) TOV21G cells (Figure 15A). In contrast, ES-2 cells had little effect as both the control and the treated cells had a PCC value of 0.69 (Figure 15B). In addition, the filamentous actin (F-actin) organization was investigated via phalloidin stain. In Figure 15C, the F-actin in the control for TOV21G showed filamentous extensions, whereas in the treated cells, the F-actin seemed to aggregate within the cell, rounding the edges.

Moreover, TOV21G cells were treated with 10μM ζ-Stat and seeded into 96 well transwell plates. After 24hrs of treatment, the cells were fixed and stained to determine the effects ζ-Stat on invasion and migration. Our data showed that ζ-Stat drastically decreased invasion and migration when compared to the control (Figure 7A). After the images were quantified, the data revealed that the decrease in invasion and migration was statistically significant (Figure 7B; invasion p value 0.002826, t = 4.859; migration p value < 0.001, t = 6.1887). To further illustrate this point, Rac1 activation was investigated utilizing a GST pull down method. The negative and positive control display how well the assay data fits the intended model (p value < 0.001, t = 7.675). Compared to the sample control, the amount of activated Rac1 pulled down from treated samples was only 37% (p value 0.044, t = -3.044; Figure 7D). These data reinforce the theory that ζ-Stat decreases the invasion and migration of CCOC through a decreased activation of Rac1.
Figure 14. Effects of ζ-Stat on Ect2 localization and RhoA protein and genomic expression. (A) TOV21G cells were seeded on a four-chamber slide and treated for 72 hrs with 10μM ζ-Stat. The slide was fixed with formaldehyde and immunostained with anti-Ect2, phalloidin dye, and DAPI. Original magnification is 40X and scale bar is 20μm. (B) Ect2 is visualized in the filamentous extension of the TOV21G cell. (C) A Western Blot for TOV21G and ES-2 cells treated for 72 hrs with vehicle control (DMSO) and 3μM ζ-Stat is represented with an immunoblot that was probed for PKC-ζ and RhoA (N=3). (D) TOV21G and ES-2 cells were treated for 72 hrs with 3μM ζ-Stat and harvested with RNAbee for mRNA semi-quantification. Endpoint PCR was run with RhoA and GAPDH primers (mRNA control) (N=3).
Figure 15. Immunofluorescent staining of Ect2. (A) TOV21G and ES-2 cells were probed with anti-Ect2, phalloidin dye and DAPI. The Pearson’s correlation coefficient is represented in white at the bottom left corner of the merged images. Original magnification is 40X and scale bar is 20μm. (B) A visualization of the organized filamentous actin in TOV21G cells treated with vehicle control (DMSO). (C) A visualization of Ect2 localization in TOV21G cells treated with 10μM ζ-Stat for 72 hrs.
Figure 16. PKC-ζ regulates invasion and migration of TOV21G ovarian cancer. (A) TOV21G cells were grown, serum starved for 48hrs, and placed in the upper chamber of a transwell plate coated with 0.1x BME and serum (10%) containing media was placed in the lower chamber as a chemo attractant. Following treatment with 10μM ζ-stat for 24hrs, cells that invaded through BME and migrated into the lower chamber were stained with crystal violet and observed under microscope. Original magnification is 10x and scale bar represents 1mm. (B) Invasion (N=4) and (C) migration (N=4) fixed and stained cells on the lower chamber were quantified using ImageJ FIJI software, average and plotted. Standard deviation error bars are represented. (** p < 0.01, *** p < 0.001). (D) Activated Rac1 (GTP bound) was pulled down using GST tagged PAK-PBD and analyzed by Western Blot. Densitometry of activated Rac1 bands were plotted (N=4). Standard deviation error bars are represented.

2.3.6 Analysis of ζ-Stat in TOV21G Tumor Xenografts

To determine the effects of ζ-Stat in-vivo, we injected athymic nude female mice with TOV21G cells and sequentially treated mice for 35 days. At the endpoint of the experiment, the tumors were harvested, and the blood serum was screened for enzymes associated with kidney and liver failure, as well as glucose levels for screening diabetes. Our data exhibited statistically significant changes in tumor volume between vehicle control and treated mice (Figure 17A) starting on day 14 (p value 0.006343, t = 3.4389) up until day 35 (p value 0.001136, t = 4.4827).
Results also demonstrated that ζ-Stat decreased tumor growth by more than 50% by the endpoint of the experiment (Figure 17B-C). The treatments did not lower the mouse population’s body weight (Figure 17D) and did not have a significant effect on the enzyme panel (Figure 17E-F). These preliminary results suggest that ζ-Stat can be used for the treatment of CCOC and does not cause short-term toxicity.

Figure 17. Effects of ζ-Stat on *in-vivo* xenografts with athymic nude mouse models. (A) Six vehicle control mice and six treatment mice were injected with 1 x 10^6 cells/per mouse flank in 0.2mL of media of TOV21G CCOC cells. Vehicle control mice tumors were treated with 1x DPBS and treatment mice were treated with 20mg/kg of ζ-Stat. Tumors were measured every other day and plotted. An unpaired student T test was performed, and standard deviation error bars are presented (* p < 0.05, ** p < 0.01). (B) Tumors were harvested after 35 days of treatment and a picture was taken. (C) Endpoint treated tumor volumes were graphed as a percent of the control (D) Mouse weight was recorded once per week. An unpaired student T test was performed, and standard deviation error bars are presented (no significant difference observed). (E-F) Endpoint blood serum was taken and screened for glucose (GLU), blood urea nitrogen (BUN), alanine aminotransferase (ALT), aspartate aminotransferase (AST) and alkaline phosphatase levels for toxicity. A unpaired student T test was performed, and standard deviation error bars are presented (no significant difference observed).
2.3.7 ζ-Stat Interrupts the PKC-ζ/Ect2 via PKC-ζ Protein Decrease

Our predicted pathway models that PKC-ζ scaffolds Ect2 to the cellular membrane (Figure 18). This mis-localization of Ect2 permits the wild-type Ect2 more access to Rac1 and therefore increases its activation. Upon increased Rac1 activation, CCOC invasion is increased. PKC-ζ protein level decrease via ζ-Stat, releases Ect2 from the membrane scaffold, and re-localizes Ect2 to the nucleus, limiting its access to cytosolic Rac1 and decreasing Rac1 activation.

Figure 18. ζ-Stat decreases the amount of PKC-ζ, therefore decreasing the scaffold of Ect2 in the cytoplasm. The hypothesized pathway illustrates the scaffolding of Ect2 via PKC-ζ in the cytoplasm. Treatment with ζ-Stat decreases PKC-ζ expression thereby there is less PKC-ζ to bind Ect2, allowing Ect2 to re-localize to the nucleus. This decreases Ect2 cellular access to Rac1, reducing Rac1 activation and cell migration.
2.4 Discussion

In this study, we discovered that the small molecule inhibitor, ζ-Stat, is a prospective drug candidate to investigate as a novel potential treatment for CCOC. We also investigated the PKC-ζ/Ect2/Rac1 activation pathway and found that ζ-Stat decreases the invasive behavior of CCOC by decreasing cytosolic Ect2 and Rac1 activation.

2.4.1 Targeting aPKCs in CCOC

To understand the therapeutic potential of a protein target, there must be an appreciation of the underlying genetic abnormalities specific to the cancer type. The TP53 gene (tumor suppressor p53) is the most commonly mutated gene in all ovarian cancers and is especially a prognostic marker of HGSOC (264). In contrast, CCOC typically has a wild type TP53 and mutations in the tumor suppressor ARID1A (265). Although PRKCZ and PRKCI are not in the top mutated genes in CCOC, PIK3CA is mutated (~30%) and is located on the third chromosome’s long arm. Interestingly, PIK3CA (3q26.32), ECT2 (3q26.31) and PRKCI (3q26.2) are all located on the long arm of chromosome three and in proximity. The ARID1A (1p36.11) is also located up stream of PRKCZ (1p36.33) on the short arm of the first chromosome. It has been noted that mutations and deficiencies in ARID1A have been shown to sensitize cancers to PARP and PI3K inhibitors (266–268).

Previous literature has suggested that the atypical PKCs should be the focus of targeted treatment (210). One explanation for this is that the atypical PKC isoforms have been linked to signaling pathways needed for cancer survival and growth. A study performed by Yao et al. illustrated the dramatic changes to prostate cells malignancy upon PKC-ζ silencing (269). In our study, when CCOC (TOV21G and ES-2) cells grew rapidly and were cell cycle arrested (serum starved), the aPKCs were present in both conditions. However, the expressions of PKC-ζ and
PKC-ι were found to be very low in the normal endometrial stromal cells. These findings may indicate that CCOC cells have a reliance on aPKC overexpression for cell viability. An interesting complication in other types of CCOC is that ζ-Stat does not specifically limit expression of PKC-ζ but also effects downstream targets in the PKC-ζ/Ect2/RhoA pathway. As shown in Figure 2, ζ-Stat had a negligible effect on PKC-ζ expression in ES-2 cells, however in Figure 5, there was a decrease in RhoA protein and mRNA expression. This leads to the conclusion that ζ-Stat has generalized effect on the pathway dependent on cell type. Furthermore, these data support that PKC-ζ/Ect2/RhoA pathway contains relevant targets for CCOC due to the lack of overexpression in normal tissue and the overexpression in cancerous cell lines.

Equally important, a previous study showed that the knock down of PKC-ζ using siRNA decreased the expression of RhoA and Rac1(270). Furthermore, this study illustrated that the knock down of PKC-ζ decreased the invasive behavior of breast cancer cells by more than 40%. Our data supports previous data and reiterates the therapeutic potential for targeting PKC-ζ in CCOC.

### 2.4.2 aPKCs are Involved in the Localization of Ect2

The mis-localization and overexpression of Ect2 has been linked to aPKCs and Ect2 dependent malignant transformation (271). Cytoplasmic Ect2 has more access to Rho GTPases and increases the protein family’s activation. Liu et al. identified Ect2 as an activator of the Par6/Par3/aPKC polarity complex and further showed that Ect2 stimulated PKC-ζ activity (272). Moreover, the oncogenic activity of Ect2 was shown to be regulated by aPKC via the phosphorylation of the Thr 328 site and a mutation in this site (T328A) rendered the Ect2 unable to interact with Par6/aPKC complex or activate Rac1 (273). Upon treatment with ζ-Stat, the decrease of PKC-ζ disturbs the localization of Ect2 in TOV21G cells. This specific effect is not
seen in ES-2 as correlations are equal before and after ζ-Stat treatment. Therefore, it does not seem that PKC-ζ is not the specific target but more of the generalized PKC-ζ/Ect2/RhoA pathway.

2.4.3 ζ-Stat has Therapeutic Potential in CCOC

A common treatment regime for ovarian cancer patients typically involves chemotherapy (paxitaxol), PARP (poly ADP ribose polymerase) inhibitors and bevacizumab (targets angiogenesis) (217). However, concerns for current treatments still involve toxicity, drug-resistance, reoccurrence and the side effects to the patients (274). Our data showed that mice treated with ζ-Stat did not have significant side effects when compared to mice treated with the vehicle control. In particular, the mice did not have significant fluctuations in body weight, or differences in the enzyme panel which screened for liver and kidney damage. Further, the viability of normal endometrial stromal cells did not change significantly upon treatment of 10μM ζ-Stat. All these data support the prospect that this compound may be a less toxic alternative maintenance drug for CCOC.

The presence of PKC-ζ and PKC-ι in proliferation, invasion and migration make this protein a unique target for therapies. In the previous literature, ζ-Stat was shown to decrease the invasion and migration of melanoma cells and increase apoptosis (225) It has also been suggested the ζ-Stat is selective to PKC-ζ in melanoma and colorectal cell lines (222,225). In support of the previous data, our data showed that ζ-Stat had a dramatic effect on CCOC invasion and migration.

However, new evidence supports the theory that ζ-Stat does not inhibit kinase activity of PKC-ζ through the ATP binding region. Our computational and in-vitro data advocates that ζ-Stat decreases the expression of signaling proteins (PKC-ζ and RhoA). This phenotypic effect
consistently modifies kinase interaction networks which results in decreases in cellular viability, invasion and tumor growth (222,225). The effects that ζ-Stat elicited was decreased PKC-ζ protein, re-localized Ect2 to the nucleus in TOV21G cells (ARID1A mutant) and decreased the presence of activated Rac1. ζ-Stat specifically decreased the protein expression of PKC-ζ in comparison to PKC-ι, however this mechanism is still unknown. The data suggests that ζ-Stat may be generating an epigenetic effect, which in turn regulates the expression of these proteins. It has been suggested that using aPKC inhibitors can epigenetically regulate the expression of aPKCs through transcription factors, such as FOX01 (225).

Additionally, according to the computational data, ICA-1S exhibited poor binding with the PKC-ι and PKC-ζ ATP site and had a predicted moderate affinity with a possible PKC-ι allosteric site. This fits well with previously performed MBP phosphorylation experiments (225). Some expected kinase activity with mid-range binding promiscuity is anticipated for this molecule. However, this conclusion is contradicted by some evidence from the experiment, therefore it could simply be that the generated PKC-ζ model is insufficient for correlating physical data. ζ-Stat exhibits extremely poor binding with the ATP site and PKC-ι allosteric site. With no expected kinase activity and low binding promiscuity, any inhibition seen using ζ-Stat is likely specific for unique binding pockets.

To further the exploration of ζ-Stat being a potential therapeutic for CCOC, a mechanistic understanding of the direct protein binding is required for PKC-ζ and PKC-ι.
2.5 Acknowledgments

The authors would like to acknowledge DeVon DeLoach in the Department of Comparative Medicine at the University of South Florida for aiding with running our serum enzyme panel and Dr. Emily Miedel, VMD, DACLAM, Associate Director of Comparative Medicine at the University of South Florida for her expert feedback and analysis.

2.5.1 Funding

We acknowledge the generous financial contribution from the Celma Mastry Foundation.

2.5.2 A Note to Reader

This work is currently under review in the journal of Frontiers in Oncology, Women's Cancer section, and is due to be published after the submission of this work.

2.6 Availability of Data and Materials

The datasets used and/or analyzed during the current study are available from the corresponding authors on reasonable request. The dataset generated and/or analyzed during the current study are available from COSMIC repository, www.COSMIC.Sanger.ac.UK.

2.7 Conflict of Interest

The authors declare that this study received funding from the Celma Mastry Ovarian Cancer Foundation. The funder was not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.
2.8 Author Contributions

TS contributions include cell culture, analysis of somatic gene mutations, cell viability assays, cell lysate collection, Western Blot analysis, semi-quantitative endpoint PCR, Rac1 activation assay, fluorescent microscopy, in-vivo experiments, and the statistical analysis of Western Blots, cell viability, particle counts for invasion and migration, tumor volume, mouse body weights, and enzyme panel. In addition, TS contributed to writing and experimental design. RM contributed by providing molecular dynamics simulations, homology modeling, molecular docking, virtual target screening, site screening, computational and statistical analysis, and writing support. RP contributed by assisting in-vivo experimentation. SMI contributed by performing the invasion and migration assay. RB contributed by performing Western blots. MD contributions to the paper include writing-review, editing, resources, supervision and funding acquisition. The authors read and approved the final manuscript.
2.9 Supplemental Figures

Figure 19. Raw Western blot images for TOV21G and SHT290. **Figure 7A.** (A) Western blot of PKC-ι in TOV21G and SHT290 cells at 50% and serum free (SF). (B) Western blot of PKC-ζ in TOV21G and SHT290 cells at 50% and serum free (SF). (C) Western blot of β-actin in TOV21G and SHT290 cells at 50% and serum free (SF). Rows 3 and 4 were used for the TOV21G image and rows 7 and 8 were used for the SHT290 image.
**Figure 20.** Raw Western blot images for ES-2 Figure 7A. (A) Western blot of PKC-ι in ES-2 cells at 50% and serum free (SF). Bands are loaded in the opposite order. (B) Western blot of PKC-ζ in ES-2 cells at 50% and serum free (SF). Bands are loaded in the opposite order. (C) Western blot of β-actin in ES-2 cells at 50% and serum free (SF). Bands are loaded in the opposite order.

**Figure 21.** Raw Western blot images for TOV21G Figure 7B. (A) Western blot of PKC-ι and its corresponding β-actin in TOV21G cells. (B) Western blot of PKC-ζ and its corresponding β-actin in TOV21G cells.
Figure 22. Raw Western blot images for ES-2 Figure 7B. (A) Western blot of PKC-ι and its corresponding β-actin in ES-2 cells. (B) Western blot of PKC-ζ and its corresponding β-actin in ES-2 cells.

Figure 23. Raw Western blot images for TOV21G Rac1 PD utilizing GST PAK-PBD Figure 16D. (A) Western blot of Rac1 exposure time of 1 minute. Rows 1 and 2 are the positive and negative control, respectively. (B) Western blot of Rac1 exposure time of 2 minutes. The positive control is saturated (marked in pink). Rows 5 and 6 are the vehicle control and 10μM treatment for TOV21G cells.
Figure 24. Raw Western blot images for TOV21G RhoA and PKC-ζ protein decrease (A) Western blot of RhoA in TOV21G cells with vehicle control (Row 2) and 3μM treatment (Row 3). (B) Western blot of PKC-ζ in TOV21G cells with vehicle control (Row 2) and 3μM treatment (Row 3). (C) Western blot of β-actin in TOV21G cells with vehicle control (Row 2) and 3μM treatment (Row 3).
3 Ligand-mediated protein degradation reveals functional conservation among sequence variants of the CUL4-type E3 ligase substrate receptor cereblon

3.1 Introduction

Cereblon (CRBN) binds to thalidomide and other immunomodulatory drugs, including lenalidomide (Len) and pomalidomide (Pom), and is one of many DDB1 and CUL4-associated factors (DCAFs) that target specific protein substrates for ubiquitylation and proteasome-mediated degradation by the DDB1–CUL4A–Roc1–RBX1 E3 ubiquitin ligase complex (Figure 33A) (275–277). Despite their teratogenic properties (278), immunomodulatory compounds have antineoplastic activity in multiple myeloma (279), myelodysplastic syndrome (MDS) associated with a somatically acquired deletion in chromosome 5 (del(5q) MDS) (280), and B cell malignancies (281–283) based on the substrates that are selected for proteosomal destruction by the bridging actions of these small molecules (Figure 33A) (275,276,284–287). Len, Pom, and a newer derivative CC-122 (avadomide) (288) also potentiate the activation of T cells. Although this function has been studied less thoroughly, it is thought that the drug-induced proteasome-mediated degradation of transcriptional repressors of T cells, Ikaros and Aiolos, may be necessary for this response in addition to their defined role in antineoplastic activity (Figure 33A) (289–291). Unlike human cells, mouse cells are resistant to these compounds, including antiproliferative multiple myeloma (292) and thalidomide-associated teratogenicity (293). Whereas the overall amino acid sequence of mouse CRBN is highly conserved (Figure 33B), and it forms the DDB1 interaction (277), minor species–related sequence variations in the

---

1 This work has been previously published in the Journal of Biological Chemistry (294).
thalidomide-binding domain (TBD) of CRBN are thought to lead to nonconserved drug binding or altered E3 ligase recruitment functions. A single amino acid substitution, Val-388, which is changed to isoleucine in mouse (Ile-391), has been reported to render mouse CRBN unable to degrade Ikaros, Aiolos, and CK1α binding through a β-hairpin–loop motif that is recognized only when CRBN is complexed with the immunomodulatory drugs (295). Therefore, a dysfunctional drug-induced substrate requirement possibly mediates drug resistance (295). Thus far, conservation of mouse CRBN's E3 ubiquitin–ligating function has not been definitely shown due to this loss of function, as the native ligands that are regulated through this pathway are poorly defined. Moreover, drug-binding affinity due to the V388I variant has not been studied in detail. A better understanding of this defect is important for future drug discovery efforts aimed at controlling intracellular protein degradation and understanding CRBN's endogenous role as an E3 ubiquitin ligase substrate receptor.

Here, we focused our analysis on nonprimate CRBN. Using drug binding assays to several CRBN sequence variants, we investigate whether these structural changes are likely to result in functional inactivation. We focused on CRBN's substrate recruiting function in mouse and human T cells, as this is poorly studied and important for defining applications in immune therapy. Both the Bradner and Crews laboratories conjugated JQ1, an established chemical inhibitor of bromodomain and extraterminal domain (BET) family members, to a thalidomide analog to promote the CRBN-dependent degradation of BRD4 (296,297) Linkage of the BET-binding ligand (JQ1 or OTX-015) to an IMiD redirects CRBN-E3 ligase activity toward BRD4, allowing for analysis of different substrates through a similar TBD-dependent mechanism (Figure 33). Moreover, the E3 ubiquitin protein–targeting approach, coined proteolysis targeting chimera (PROTAC) (298–300), may be broadly applied clinically and experimentally to study
the function of proteins that are difficult to target and in cells that are insensitive to genome editing techniques. Here, we use a PROTAC probe to investigate cellular engagement of an immunomodulatory drug with mouse CRBN. Collectively, our findings are important to understand sequence variants of CRBN and provide the first investigation of the E3 ubiquitin ligase substrate-conjugating function of CRBN in a nonprimate vertebrate model.

3.2 Results

3.2.1 Resistance of mouse cells to immunomodulatory compounds.

Comparative analyses of CRBN sequences from representative vertebrate species revealed that Ile-391 is conserved among many nonprimate mammals (mouse, rat, dog, manatee, and opossum), birds (chicken and zebra finch), reptiles (alligator), amphibians (Xenopus), bony fish (zebrafish and spotted gar), and cartilaginous fish (whale shark) (Figure 33B) (301). Interestingly, both chicken and zebrafish are susceptible to thalidomide-induced teratogenicity despite expressing the Ile-391 variant (277). In contrast, Val-388 is present in all examined primate sequences (human, chimpanzee, gorilla, orangutan, gibbon, macaque, baboon, green monkey, squirrel monkey, marmoset, tarsier, and Sunda flying lemur) with the exception of the gray mouse lemur sequence, which encodes Ile in this position (Figure 33B). Therefore, Val-388 is a recently derived feature present in most primates.

Immunomodulatory drug–induced ubiquitin-mediated degradation of Ikaros and Aiolos (encoded by IKZF1 and IKZF3, respectively) (275,302) appears sufficient to augment IL-2 production by T cells (303) in the absence and presence of anti-CD28 co-stimulation (289,304,305). In Len-treated human T cells stimulated with anti-CD3ε antibody to cross-link
the T-cell receptor (TCR), levels of IL2 mRNA (Figure 34A) and protein (Figure 34B) were significantly increased relative to DMSO (vehicle)-treated cells. Comparing purified human (Figure 25A) and mouse T cells (Figure 25B) pretreated with vehicle or 10 μM Len, only human T cells displayed the expected Len-induced increase in IL-2 when stimulated in the absence (Figure 25, A and B) or presence (Figure 34, C and D) of anti-CD28 antibody using doses of anti-CD3ε that ranged from 0.01 to 10 μM.

Thalidomide, Len, and Pom's antiproliferative effects in multiple myeloma cell lines also reportedly differ based on the presence of mouse CRBN (292,295,306). In U266 (Table 11), H929 (Table 11), and MM1.S multiple myeloma cell lines (Figure 25C and Table 11), the antiproliferative effects (IC$_{50}$ values) of Pom ranged from 0.05 to 0.51 μM, and for Len, they ranged from 1.5 to 10 μM. In contrast, the mouse multiple myeloma cell line 5TGM1 was resistant to immunomodulatory drugs (Figure 25D and Table 11). Further, Ikaros protein expression was unaffected in primary mouse T cells versus almost completely depleted in human cells (Figure 25E), as predicted from previous structural and functional studies in multiple myeloma cell lines and in the Ba/F3 mouse lymphoma cell line (287,295).
Figure 25. Mouse cells are resistant to immunomodulatory drugs. Shown are T cells purified from healthy donor peripheral blood mononuclear cells (A) or from mouse spleens of C57BL/6j mice (B) and stimulated in the presence of increasing concentrations of anti-CD3ε antibody in the presence of 10 μM Len or vehicle control (DMSO) without anti-CD28. IL-2 production was determined from the culture supernatant by ELISA. The human MM1.S (C) or mouse 5TGM1 (D) multiple myeloma cells were cultured for 7 days with increasing concentrations of Len, Pom, and vehicle (DMSO). Percentage relative cell viability is shown. E, Western blot analysis of Ikaros, CRBN, and β-actin (loading control) in human and mouse T cells stimulated with anti-CD3ε 5 μg/ml + 1 μg/ml anti-CD28 antibody for 24 h with DMSO and 10 and 20 μM lenalidomide and pomalidomide. Results are representative of three independent experiments. Statistical analysis was conducted using ANOVA, followed by Dunnett's multiple comparison test. *, p < 0.05; ***, p < 0.001 (see also Figure 34). Error bars, S.D.
Figure 26. Structure of cereblon is conserved across different species. A, ribbon overlays of human (purple; PDB code 4TZ4) and chicken (blue; PDB code 4C12) X-ray crystal structures and post-MD structure of mouse (yellow; PDB code 4TZ4) cereblon. Human immunomodulatory drug–binding site residues His-378, Trp-380, Trp-386, and Phe-402 and ligands thalidomide (yellow), lenalidomide (green) and pomalidomide (blue) are shown for reference. B, superposition of ligand poses of lenalidomide for the post-MD equilibrated systems of the CRBN thalidomide-binding site after IFD for hCRBN (red), hmCRBN (green), and gCRBN (orange), shown with the post-MD equilibrated protein structure of hCRBN (red) for reference. C, overlays of human and mouse cereblon show nonconserved residues Ile-391 for mCRBN (human Val-388), Val-380 for mCRBN (human Glu-377) (mouse shown in italic type) (see also Figure 43 – Figure 45).
3.2.2 Thalidomide-binding domain of CRBN has a conserved immunomodulatory compound–binding motif

Next, the sequence variants of CRBN were studied in more detail. Based on crystal structures, immunomodulatory compounds bind to a conserved pocket within the C-terminal TBD of CRBN (Figure 33 and Figure 26A) (307,308). These interactions are governed by hydrogen bonding, aromatic quadrupole, and van der Waals interactions. Analysis of the X-ray crystal structures of CRBN (human (hCRBN), mouse (mCRBN), and chicken (gCRBN)) in complex with thalidomide, Len, and Pom, respectively (Figure 26A), shows negligible variations of root mean square deviation (RMSD) between the inhibitor poses. Thus, an in-depth theoretical investigation of the molecular binding mechanics of immunomodulatory compounds in complex with CRBN was conducted to explore possible differences in drug interactions between mouse and human CRBN caused by induced fit or protein flexibility.

For molecular dynamics (MD), simulations of the crystal structures of hCRBN (Protein Data Bank (PDB) code 4TZ4) (307) and gCRBN (PDB codes 4CI1, 4CI2, and 4CI3) (307) in complex with DDB1 were used. The crystal structure of mCRBN (PDB codes 4TZC and 4TZU) (308) is monomeric and truncated to contain only the TBD (108 defined residues compared with 380 defined residues for hCRBN and gCRBN) and is not suitable for computational modeling. The amino acid differences in mouse CRBN (when compared with human CRBN) include C366S, E377V (which may be specific to rodents), and V388I (Figure 26C). Therefore, to further increase sampling and determine any structural dependence on these residues, the original hCRBN sequence of the equilibrated representative structure was mutated to the mouse sequence. The mCRBN analog and hmCRBN hybrid developed from the hCRBN system is capable of reproducing the crystal ligand poses of mCRBN (PDB codes 4TZC and 4TZU) with
minimal conformational deviation (Figure 26C) and was used for modeling purposes. Binding modes do not appear to differ between models and compounds, as there are no significant differences in RMSD calculations between X-ray crystal binding poses (Table 12) and post-MD equilibrated models (Figures 2, B and C). Induced-fit docking (IFD) also predicts no observable difference in binding affinity between models and compounds (Table 12). All poses are within 1.8 Å RMSD, which is the expected threshold for the IFD protocol (124).

Although Val-388 of hCRBN recruits Ikaros, Aiolos, and CK1α upon immunomodulatory compound binding (295), the side chain is >6 Å away from the immunomodulatory drug, and it is thus unlikely to alter binding affinity to the drugs. The second distinct amino acid is Glu-377, which in the mouse is Val-380 and could establish a weak hydrogen bond with Len's amino group. However, hydrogen-bond analysis (Figure 38) of MD simulations suggests that minimal interaction occurs between this residue and bound immunomodulatory drugs, mainly due to the backbone dihedral strain tending to force the charged carboxyl moiety away from the binding site.

3.2.3 Immunomodulatory compound binding is conserved in CRBN sequence variants

Amino acids in mouse CRBN–TBD at Val-380 (equivalent to human Glu-377) and Ile-391 (equivalent to human Val-388) (Figure 27, A–C) appear to have no relevance in the structure or corresponding immunomodulatory drug–binding interaction based on theoretical modeling. To test the effects of these two nonconserved amino acids on binding affinity, the recombinant human TBD motif (residues 319–425) was expressed in Escherichia coli (Figure 39A) and mutated to the mouse variants (Figure 27, A–C). Immunomodulatory drug binding
was then analyzed using two distinct assays, an intrinsic tryptophan fluorescence assay (IF) (Figure 27, D–G) and isothermal titration calorimetry (ITC) (Figure 27, H–K). The C366S amino acid mutation was not studied in binding assays, as it is more than 20 Å away from the immunomodulatory drug–binding pocket. The TBD is structurally stabilized by four cysteine residues (Cys-323, Cys-326, Cys-391, and Cys-394) that coordinate a single zinc ion (307,308), located ~18 Å from the drug-interacting site. To gain insights into the role of zinc, mutations in the CXXC domain of the TBD were also generated. Mutating any of the cysteine residues resulted in insoluble protein that aggregated in inclusion bodies (Figure 39B). This is indicative of misfolding due to loss of Zn$^{2+}$ ion coordination. To rule out improper folding or destabilization, a zincon assay (309) was performed on purified protein of all expressed recombinant CRBN–TBD proteins. These analyses revealed a 1:1 stoichiometric ratio of Zn$^{2+}$ bound to the TBD recombinant protein (Figure 39, C–F). Moreover, protein secondary structure consistent with proper folding was also evident using CD (data not shown). Both IF (Figure 27, D–G) and ITC (Figure 27 (H–K) and Figure 40) analyses demonstrated similar $K_D$ values at equilibrium for thalidomide and the immunomodulatory compounds tested in binding to WT, E377V, V388I, and E377V/V388I hmCRBN–TBD (Table 10) with no binding observed by phthalimide (Figure 40, I–L), used as a negative control.
Figure 27. Human and mouse CRBN binds IMiDs with similar affinities. A, sequence alignment of human CRBN and human-to-mouse mutations. Mutations introduced to convert human to mouse are highlighted in red. B, IMiD interaction in the hydrophobic binding pocket. C, lenalidomide (green) interacts with the TBD site (gray) through hydrogen bonds (dashed black lines) with backbone residues His-378, Ser-379, and Trp-380 van der Waals interactions (dashed green lines) that occur with the side chains of Trp-380, Trp-386, Trp-400, and Phe-402 (mouse: Trp-383, Trp-389, Trp-403, and Phe-405). The two residues differing between the human and the mouse proteins are highlighted in cyan. Shown are titration of human TBD WT (D), E377V (E), V388I (F), and E377V/V388I (hmCRBN–TBD) (G) to lenalidomide (red), pomalidomide (green), thalidomide (blue), and phthalimide (black) by intrinsic tryptophan fluorescence assay. K_D values were calculated based on the magnitude of fluorescence differences (1 – F/F0). H–K, isothermal titration calorimetry saturation curve using Len for human TBD and mutants (see also Figures S3–S5 and Table 10). Error bars, S.D.
Table 10. Binding affinity (K$_D$, [μM]) of immunomodulatory compounds to WT and mutant human CRBN-TBD determined using ITC and fluorescence intensity assay.

<table>
<thead>
<tr>
<th>TBD</th>
<th>Assay</th>
<th>Lenalidomide</th>
<th>Pomalidomide</th>
<th>Thalidomide</th>
<th>Phthalimide</th>
<th>dBET1</th>
</tr>
</thead>
<tbody>
<tr>
<td>hCRBN</td>
<td>ITC</td>
<td>11 ± 3</td>
<td>16 ± 4</td>
<td>65 ± 40</td>
<td>NB</td>
<td>NP</td>
</tr>
<tr>
<td></td>
<td>FI</td>
<td>13 ± 2</td>
<td>16 ± 3</td>
<td>38 ± 12</td>
<td>NB</td>
<td>16 ± 2</td>
</tr>
<tr>
<td>hCRBN-H378A</td>
<td>ITC</td>
<td>22 ± 3</td>
<td>35 ± 8</td>
<td>37.4 ± 10</td>
<td>NB</td>
<td>NP</td>
</tr>
<tr>
<td></td>
<td>FI</td>
<td>27 ± 4</td>
<td>22 ± 6</td>
<td>74 ± 20</td>
<td>NB</td>
<td>NP</td>
</tr>
<tr>
<td>hCRBN-W380A</td>
<td>ITC</td>
<td>NB</td>
<td>NB</td>
<td>NB</td>
<td>NB</td>
<td>NP</td>
</tr>
<tr>
<td></td>
<td>FI</td>
<td>NB</td>
<td>NB</td>
<td>NB</td>
<td>NB</td>
<td>NP</td>
</tr>
<tr>
<td>hCRBN-E377V</td>
<td>ITC</td>
<td>9.6 ± 2</td>
<td>35 ± 9</td>
<td>97 ± 100</td>
<td>NB</td>
<td>NP</td>
</tr>
<tr>
<td></td>
<td>FI</td>
<td>22 ± 2</td>
<td>30 ± 7</td>
<td>36 ± 8</td>
<td>NB</td>
<td>NP</td>
</tr>
<tr>
<td>hCRBN-V388I</td>
<td>ITC</td>
<td>10 ± 1</td>
<td>28 ± 7</td>
<td>43 ± 20</td>
<td>NB</td>
<td>NP</td>
</tr>
<tr>
<td></td>
<td>FI</td>
<td>14 ± 4</td>
<td>16 ± 4</td>
<td>39 ± 1</td>
<td>NB</td>
<td>13 ± 2</td>
</tr>
<tr>
<td>hmCRBN-E377V/V388I</td>
<td>ITC</td>
<td>14 ± 2</td>
<td>8 ± 2</td>
<td>34 ± 2</td>
<td>NB</td>
<td>NP</td>
</tr>
<tr>
<td></td>
<td>FI</td>
<td>13 ± 3</td>
<td>19 ± 4</td>
<td>30 ± 8</td>
<td>NB</td>
<td>NP</td>
</tr>
</tbody>
</table>

Data represent mean ± S.D. FI, fluorescence intensity; NB, non-binding; NP, not performed.

To assess the impact of binding pocket residues, Ala mutations of two residues were generated (Table 10) (281,301,302) to test the impact of hydrogen bond formation and hydrophobic interaction with Trp-380. The W380A mutation completely abolished ligand interactions. Although His-378 forms two hydrogen bonds with the glutarimide ring (Figure 27C), mutating this residue to Ala showed only minor impact on binding. Pom and Len affinities were reduced, while thalidomide displayed increased affinities. This suggests that the backbone carbonyl of H378A may retain the hydrogen bond interaction with the NH group of the glutarimide ring. To further probe the role of His-378 side chain in immunomodulatory drug binding, we conducted a pH dependence study to measure the binding affinity of Len to CRBN–
TBD by ITC. The $K_D$ values measured at pH 4.5, 5.5, 6.5, and 7.5 are $21.4 \pm 3$, $23.7 \pm 8$, $23.8 \pm 7$, and $11.3 \pm 2 \mu M$, respectively. Some impact was seen at pH 7.5, indicating deprotonation of the His-378 imidazole group has marginal impact on the immunomodulatory drug binding to the TBD. However, the protein microenvironment influence on the histidine glutarimide ring can easily overcome small deviations of 1 or 2 bulk pH. Thus testing affinities under the aforementioned conditions are likely insufficient.

### 3.2.4 N-terminal stabilization of CRBN–immunomodulatory compound interactions

Finally, we compared binding affinities of Len to CRBN–TBD and full-length CRBN–DDB1 protein complex using ITC to assess the impact of residues outside of the TBD. The full-length CRBN–DDB1 complex displayed a $K_D$ value of $0.64 \mu M \pm 0.24 \mu M$ (pH 7.0) (Figure 28A). This affinity is similar to published data using a fluorescence polarization-based assay(307). Moreover, these results are consistent with a single binding site within the protein complex. To gain more insight into immunomodulatory compound binding to CRBN–TBD, we synthesized N-methyl-Len as a negative control, where the N-methyl group of the glutarimide ring is predicted to cause steric hindrance in the binding pocket (Figure 41) and lacks the key hydrogen bond donor to His-378. As expected, N-methyl-Len did not bind to either the CRBN–TBD or CRBN–DDB1 complex (Figure 28B), as measured by ITC, indicating that the complex–drug interaction is mediated predominantly by the glutarimide binding pocket of the TBD. Interestingly, the binding affinity of Len to the CRBN–TBD is about 30-fold lower than the full-length CRBN–DDB1 complex. Moreover, full-length CRBN and CRBN–TBD have endothermic and exothermic reactions, respectively. Therefore, residues in the full-length protein appear to augment protein-ligand interactions in the binding pocket.
Figure 28. Lenalidomide binds to the TBD and CRBN–DDB1 protein complex. ITC binding curves of lenalidomide (A and C) and N-methyl-lenalidomide (B and D) titrated with CRBN–DDB1 complex (A and B) and CRBN–TBD (C and D). E and F, schematic view of lenalidomide interaction in the binding pocket of full-length human cereblon; human full-length CRBN (salmon) and DDB1 (blue) with bound lenalidomide (magenta) (PDB code 5FQD) superimposed with mouse TBD-CRBN (cyan) (PDB code 4TZU). Hydrogen bonds and hydrophobic interactions are shown in red and black dashed lines, respectively (see also Figure 41). Data in figure represent mean ± S.D.
3.2.5 Acquired ubiquitin-proximity ligation in mouse cells establishes conserved ligase functions of mouse and human CRBN variants.

The substrate-recruiting function of CRBN was then investigated using dBET1, which maintains the CRBN–TBD targeting domain but switches the substrate-recruitment domain to recognize BRD4 and other JQ1-associated targets (296). Using IF (Figure 29A), the saturation binding curves for dBET1 are similar to that of Len, whereas, as expected, JQ1 alone does not bind to the CRBN–TBD. The structures of dBET1 and N-methyl-dBET1 (used as a negative control, based on the results of N-methyl-Len) are provided in Figure 29B to show the CRBN- and BET-targeting groups. Moreover, N-methyl-dBET1 confirms that this analog is interacting with similar residues in the TBD. Next, the function of dBET1 in activated human T cells was assessed, and results are shown in Figure 29 (C–G); statistical analysis is provided in Table 13. Unlike Len and Pom, which activate T cells, BRD4 and c-Myc inhibition is expected to induce cell death or functionally repress these cells, as they are critical mediators of T-cell proliferation and survival (310,311). Relative viability of activated T cells is reduced by dBET1 treatment compared with DMSO and is more active relative to treatment with increasing doses of JQ1 (Figure 29C), suggesting that growth suppression by the heterobifunctional conjugate is superior to JQ1, as shown previously in leukemia cells (296). Importantly, decreased BRD4 protein expression is only evident with dBET1 treatment, whereas both dBET1 and JQ1 suppress the activation of c-Myc protein expression by Western blot analysis (Figure 29D) and flow cytometry (Figure 29E; summary of data in Figure 29F), as expected, through BRD4 functional repression. Notably, the effect on BRD4 is reversed by incubation with N-methyl-dBET1 at doses of 0.1–10 μM. Repression of c-Myc at very high doses of N-methyl-dBET1 is consistent with activity of the JQ1-targeting molecule (Figure 29, D and F). From these results, we
conclude that human T cells respond to dBET1 through interactions that are mediated by the hydrophobic pocket of the CRBN–TBD. Finally, the suppressive effects observed with dBET1 treatment shows that the CRBN-targeting molecule in this compound is no longer activating IL-2, as shown in Len and Pom (Figure 29G).

Figure 29. Functional activity of lenalidomide, JQ1, and dBET1 in human T cells and binding affinities of JQ1, dBET1, and Len to human CRBN. A, saturation binding curves of human TBD WT titrated with lenalidomide ($K_D = 15.6 \pm 2.2$), JQ1 (no binding), dBET1 ($K_D = 26.0 \pm 2.1$),
and N-methyl-dBET1 (no binding) by fluorescence assay (see also Table 10 for a summary of all
$K_D$ values). B, structure of dBET1 and N-methyl-dBET1. C, human T cells purified from
peripheral blood mononuclear cells from healthy donors and stimulated with anti-CD3ε/CD28.
At the time of TCR stimulation, DMSO (vehicle control), the indicated concentrations of JQ1
and dBET1, 10 μM Len, and 10 μM Pom were added to the cell cultures. Cell viability was
determined after 72 h by flow cytometry using Zombie NIRT staining. D, Western blot
analysis for expression of BRD4, CRBN, c-Myc, and vinculin in human T cells stimulated with
anti-CD3ε/CD28 and treated simultaneously with vehicle (DMSO, 0.1%), JQ1, dBET1 (10 μM),
and N-methyl-dBET1 (10 μM) for 48 h. Unstimulated cells are not shown in D. E, c-Myc
expression by flow cytometry (histogram data) is shown following 48-h stimulation with anti-
CD3ε/CD28. Drug treatments and doses are indicated. Expression of c-Myc requires stimulation
based on comparison with unstimulated cells (unstim). Values represent percent positive relative
to unstimulated cells based on gate (dotted line). F, quantification of flow data shown in E
represented as percentage of DMSO. Bar height is the mean of individual values shown
representative of three independent experiments. Error bars, S.D. from an exemplary experiment.
G, unstimulated (unstim) or anti-CD3ε/CD28–stimulated T cells were treated with 10 μM Len,
10 μM Pom, and the indicated doses of JQ1 or dBET1. ELISAs were used to quantify IL-2
concentrations (pg/ml) from supernatants harvested at 48 h. Results shown are representative
of three independent experiments. Bar height, mean of individual values; error bars, S.D. from an
exemplary experiment. Results of statistical analysis are provided in Table 13.

Next, saturation binding curves of V388I-TBD mutant titrated with Len ($K_D = 15.6 \pm 2.2$
μM), dBET1 ($K_D = 26.0 \pm 2.1$ μM), and N-methyl-dBET1 (nonbinding control) show that the
mode and binding affinity of dBET1 are similar to those of the human-TBD protein (Table 10
and Figure 30A). Mouse T cells were purified from Crbn$^{+/+}$ and Crbn$^{-/-}$ mice and used to
evaluate the role of mouse CRBN in BRD4 degradation (Figure 30, B and C), activation-
induced proliferation, viability, and c-Myc and CD98 (Myc target gene) (Figure 30, D–G)
expression in response to dBET1 treatment (Figure 29B). TCR stimulation with anti-CD3ε +
anti-CD28 is required for the induction of c-myc in primary mouse T cells (data not shown).
Both JQ1 and dBET1 suppressed c-Myc in activated T cells, but the reduction in protein
expression of BRD4 was only observed after dBET1 treatment in Crbn$^{+/+}$ mouse T cells,
consistent with proteasome engagement (Figure 30B) through a CRBN-dependent mechanism.
Pomalidomide treatment failed to impact BRD4 expression and further supports redirection of CRBN's ubiquitin-conjugating function through JQ1. To assess whether dBET1 is interacting directly with the TBD of mouse CRBN in vivo, an assay was performed using mouse Crbn<sup>+/+</sup> T cells treated with Pom in the presence and absence of dBET1. Results shown in **Figure 30C** show that at high concentrations of Pom (10-fold excess), the impact of dBET1 on c-Myc and BRD4 expression was reversed, suggesting that dBET1 is directly engaging the TBD of mouse CRBN. We next examined proliferation (i.e. division index) measured by dilution of cell trace violet (CTV) using methods described previously (312), expression of a c-Myc target gene (CD98), and viability using flow cytometry (**Figure 30D**). Histograms (**Figure 30D**) and summarized results (**Figure 30, E–G**) show that the division index (based on CTV data) and CD98 expression (median fluoresce intensity (MFI)) are induced through activation (unstimulated versus DMSO) and that there is a differential response to dBET1 treatment in Crbn<sup>+/+</sup> and Crbn<sup>−/−</sup> T cells at 0.1 and 1 μM. The suppression of T cells at 10 μM is probably related to the JQ1-targeting molecule, because proliferation, c-Myc–regulated CD98 expression, and viability were suppressed independently of CRBN expression. Moreover, BRD4 degradation by dBET1 is similar in MM1.S human and 5TGM1 mouse multiple myeloma cells, confirming the conserved ubiquitin-conjugating functions of CRBN (**Figure 42**).
**Figure 30.** Functional activity of lenalidomide, pomalidomide, JQ1, and dBET1 in mouse T cells. A, saturation binding curves of human V388I mutant TBD titrated with lenalidomide (K_D = 18 ± 4), JQ1 (no binding), dBET1 (K_D = 18 ± 3), and N-methyl-dBET1 (no binding) by fluorescence assay (see also Table 10). B, mouse T cells purified from spleens of C57BL/6j mice with homozygous deletion of Crbn and WT littermates were stimulated with anti-CD3/CD28 in the presence of 0.1% DMSO (vehicle control), 10 μM Pom, 10 μM JQ1, or 10 μM dBET1. Western blots for BRD4, c-Myc, CRBN, and vinculin (loading control) expression are shown. C, Western blot analysis showing expression of BRD4, c-Myc, and vinculin from Crbn^{+/+} T cells stimulated with anti-CD3/CD28 for 12 h with the indicated doses of dBET1, Pom, or combined treatment. D, purified Crbn^{+/+} and Crbn^{−/−} mouse T cells stimulated with anti-CD3/CD28 and treated with 0.1% DMSO (vehicle control), 10 μM Len, 10 μM Pom, 10 μM JQ1, or increasing doses of dBET1 for 72 h. Histogram plots from one experiment show proliferation (determined by dilution of CTV) and CD98 expression of T cells. E, proliferation index based on CTV data was used to calculate the amount of proliferation using an algorithm available with Flowjo analysis software (Tree Star). F, expression of CD98 was calculated relative to unstimulated cells using gates shown based on mean fluorescence intensities (MFI) of T cells as shown in D. G, cell
viability was determined after 72 h by flow cytometry using Zombie NIRTM staining. Results shown are representative of three independent experiments. Bar height represents mean of individual values, and error bars represent S.D. from an exemplary experiment. Results of statistical analysis are provided in Table 14.

3.3 Discussion

CRBN was first identified in mild autosomal recessive nonsyndromic intellectual disability (313) but has poorly defined physiological functions. Interactions have been reported with the AMP-activated protein kinase α1 subunit (314), TAK1–TRAF6 (315), and CD147–MCT1 complex (316), where CRBN plays ubiquitin-independent roles in pathway regulation. The mechanistic underpinnings for induced limb deformities in chickens and zebrafish, but not in mice, appears dependent on sequence differences in mouse CRBN (295), which brings into question whether these differences in CRBN render it functionally inactive. Although reference CRBN sequences for nearly all examined primates possess Val-388, according to the Exome Aggregation Consortium, the Ile-391 variant occurs in humans at a frequency of 0.005% (rs756414303). Our data show that mutating Val-388 to Ile (V388I) does not alter binding affinity to the immunomodulatory compounds, suggesting that the contact between the immunomodulatory drug and the CRBN–TBD is maintained and potentially functionally active.

When bound to immunomodulatory drugs (Len, Pom, and CC-122), CRBN induces the destruction of three substrates, Ikaros, Aiolos, and CK1α (287), via a Val-388 interaction (275). In T cells, the IKZF-family transcription factors repress IL2 so that ubiquitin-mediated destruction may be responsible for promoting T cell activation (287,290). Our data solidify the mechanistic involvement of CRBN Ile-391 in drug resistance in mouse cell lines, as suggested previously by the finding that overexpression of hCRBN with Val-388, but not Ile-391, in Ba/F3 cells was sufficient to restore Ikaros protein destruction (287,295).
As expected, using dBET1, there was no Ikaros degradation in mouse T cells or mouse myeloma cell lines. However, we found that dBET1 impressively reduced the expression of BRD4 in both human and mouse cells. Therefore, proximity-associated ubiquitin-conjugating functions of mouse and human CRBN are confirmed using dBET1. Crbn−/− T cells demonstrated that dBET1 functions through a CRBN–TBD–dependent process. Competition with Pom further suggests that the TBD of human and mouse CRBN have similar binding modes for immunomodulatory compounds, suggesting that there is both structural and functional conservation. Thus, the endogenous E3 ubiquitin ligase activity and assembly of the CRBN–DDB1–CUL4A–containing complex are fundamentally conserved across the vertebrate lineage, including the mouse, and probably other species, expressing Ile-391. Moreover, other sequence variants present in mouse CRBN, including difference in the N terminus, fail to impact its substrate-recruiting functions when targeting BRD4. This observation, along with the high degree of CRBN conservation, indicates that selective pressure has maintained the overall structure and function in CRBN for over 400 million years.

Using ITC and fluorescence-based binding assays, we establish that the dissociation constants of thalidomide and other immunomodulatory compounds to mouse CRBN are similar to human CRBN, which is consistent with analysis of the crystal structure of human CRBN–DDB1 in complex with Len (307,308). As suggested previously (307), the W380A mutant completely abolished binding, which confirms that this is one of the key residues of the binding pocket. Trp-380 also appears to work synergistically with other binding pocket residues (Trp-386, Trp-400, and Phe-402) for ligand interaction. The $K_D$ values of immunomodulatory compounds in complex with the human TBD (amino acids 319–425) are in the micromolar range, similar to those of Caenorhabditis elegans and Magnetospirillum gryphiswaldense (317).
Based on the conserved CRBN structure, we synthesized N-methyl derivatives of Len and dBET1, which proved useful in assessing the functional contribution of this ligand interaction in vivo in mouse and human cells.

Here, we show an interesting reliance of the rigid TBD-binding pocket on N-terminal sequences, which plays a previously unappreciated role in optimal ligand binding. A conformational change in full-length CRBN–DDB1 may occur upon ligand binding, as shown by the difference in their $\Delta H$ and $\Delta S$ values (Figure 28). Notably, the CRBN–DDB1 complex and CRBN–TBD show a marked difference in Len-binding affinity by ITC. In fact, a closer inspection of the crystal structure of full-length CRBN in complex with Len illustrates that a disordered loop in the TBD consisting of residues Asn-351, Pro-352, and His-353 stabilizes the interaction of Len. Asn-351 forms a hydrogen bond to a carbonyl oxygen of Len's isoindolinone ring, and both Pro-352 and His-353 form hydrophobic interactions with the aromatic system of the immunomodulatory drug (Figure 28E). Importantly, the mouse TBD crystal structure shows poor electron density in this loop structure, suggesting that it is a highly unstructured and dynamic region. Alternatively, other residues outside the TBD could have stabilized this loop.

The side chain of Gln-100, which is located in the LON domain of CRBN, forms a weak hydrogen bond interaction to the $\epsilon^2$NH group of His-378. This in turn positions the $\delta^1$NH of His-378 as a hydrogen donor to the immunomodulatory drugs. These important structural domains should be further investigated in immunomodulatory drug discovery.

Collectively, our results suggest that PROTAC molecules and possibly other CRBN-bound compounds may adopt an active conformation that is susceptible to CRBN-directed, cullin–RING E3 ligase–mediated polyubiquitination in mouse cells. PROTAC molecules are designed to harness the CRBN-binding properties of IMiDs and initiate the degradation of
oncogenic targets. Several IMiD-based BET-targeting PROTACs, such as dBET-1, ARV-825, and BETd-260 have been developed to potently degrade BRD4 (297,318,319). Intracellular protein degradation of FK506-binding protein (FKBP12) was also achieved by a conjugate of thalidomide (296,320), but originally, methionine aminopeptidase-2 (MetAP-2) (300), estradiol, and dihydroxytestosterone were degraded by engaging the Skp1–Cullin–F-box (SCF) ubiquitin ligase through a 10-amino acid phosphopeptide derived from IκBα (298,321). PROTAC with specificity for the von Hippel–Lindau ubiquitin ligase E3 have also been developed (322,323). Recently, BRD4 and ERK1/2 degradation was induced by the interaction of two smaller precursors molecules that undergo intracellular self-assembly, which improves solubility and cellular permeability of thalidomide-containing PROTAC inhibitors (324). Importantly, our studies establish that mouse platforms can indeed be used for preclinical development of dBET1 and possibly other PROTAC-based chemical degraders that are designed to redirect CRBN's substrate-binding function toward specified endogenous proteins (296,298,300,318,324,325). Toxicology and functional testing of such agents in rodents and mouse tumor models may yield important preclinical information.

3.4 Experimental procedures

3.4.1 Animals and cell lines

Germ line Crbn-deficient mice (Crbn^−/−) were described previously (313), and gene deletion was confirmed using WT and Crbn-KO–specific primers. C57BL/6 (Crbn^+/+) mice were purchased from Jackson Laboratory (Farmington, CT) and were then bred to Crbn^−/− mice. Crbn^+/+ and Crbn^−/− littermates from Crbn^+/− intercrosses were used for our studies. Mice were
maintained and bred at the H. Lee Moffitt Cancer Center and Research Institute under a protocol approved by the institutional animal care and use committee. The human multiple myeloma cells, including U266, H929, and MM1.S, and the mouse multiple myeloma cell line 5TGM1 were generous gifts of Drs. Ken Shain and Connor Lynch (Moffitt Cancer Center, Tampa, FL). All cell lines were mycoplasma-free and sequence-verified.

3.4.2 T-cell isolation, activation, and drug treatments

Human polyclonal CD3+ T cells or CD8+ T cells were isolated from peripheral blood donations to the Southwest Florida Blood Services. Because personal identifying information is unavailable, the research was deemed nonhuman research. Human and mouse T cells were isolated from Crbn+/+ and Crbn−/− splenocytes by immunomagnetic negative selection (Miltenyi Biotec, San Diego, CA), and >95% purity was confirmed by flow cytometry. For drug treatment experiments, 12-well flat bottom plates were coated with 5 μg/ml anti-CD3 (clone HIT3a (eBioscience) or clone (145-2C11)) in 1× PBS at 37 °C for 60 min. Cells were plated at 2–4 × 10⁶ cells/well with anti-CD28 (clone CD28.2 (eBioscience) or clone 37.51 (eBioscience)). Following 12 h of activation, the cells were treated with DMSO (0.1%, Sigma-Aldrich), Len (10 μM) (Celgene, NJ), Pom (Sigma-Aldrich), and JQ1 (doses indicated) (catalog no. SML0974, Sigma-Aldrich). N-Methyl-Len, dBET1, and N-methyl-dBET1 were all synthesized at the Moffitt Cancer Center (described in the supporting material) and used at the doses indicated. After 12 h of drug treatment, cells were harvested, and protein levels were examined by Western blot analysis. For proliferation experiments using mouse T cells, 0.1–10 μg/ml anti-CD3ε (clone 145-2C11, eBioscience) was used with cells plated with and without anti-CD28 for 72 h. Cytokine expression was determined using supernatants that were harvested at 24 or 48 h and
quantified from standard curves by ELISA according to the manufacturer's protocol. Kits were purchased from eBioscience (IL-2) and R&D Systems for other cytokines. For functional analysis of T cells treated with JQ1, murine CD3⁺ T cells from Crbn⁺/+ and Crbn⁻/⁻ splenocytes and human T cells were labeled with 5–10 μM CellTrace Violet (C34557, Thermo Fisher Scientific) and activated with 5 μg/ml anti-CD3ε and 1 μg/ml anti-CD28 for 72 h in round-bottom 96-well plates. Cells were stained with CD98-PE (clone RL388, Biolegend), 7-aminoactinomycin D (BD Pharmingen), and the Zombie NIR™ fixable viability kit (catalog no. 423105, Biolegend) and analyzed on a BD LSRII flow cytometer.

### 3.4.3 Quantitative real-time PCR

Isolated T cells from Crbn⁺/+ and Crbn⁻/⁻ mice were lysed and homogenized (Qiashredder, Qiagen), and total RNA was extracted (RNeasy, Qiagen) according to the manufacturer's protocol. Complementary DNA was generated from isolated RNA (iScript cDNA synthesis kit, Bio-Rad). RNA expression was analyzed by quantitative real-time PCR using Taqman Universal PCR Master Mix for Taqman probes (Thermo Fisher Scientific) against (cDNA) c-Myc (Mm00487804_m1) and (cDNA) β2 M (Mm00437762_m1). Samples were run on an Applied Biosystems 7900 HT and Sequence Detection Systems software.

### 3.4.4 Treatment of multiple myeloma cells

Mouse and human multiple myeloma cell lines were plated at 2–4 × 10⁶ cells/well in a 12-well plate with various concentrations of Len and Pom. To confirm target degradation, the cells were treated with varying concentrations of dBET1 (0, 0.01, 0.1, 1, and 10 μM) for 12–24 h. Following drug treatment, protein levels were examined by Western blot analysis relative to
vinculin or β-actin to normalize for protein expression. For proliferation studies, 1–2 × 10^4 cells/well were seeded in a 96-well plate and were treated with using the CCK8 (cell-counting-8) kit (Dojindo, Rockville, MD) according to the manufacturer's protocol.

3.4.5 General chemistry information

All reagents were purchased from commercial suppliers and used without further purification (except where mentioned otherwise). 1H NMR spectra were recorded on an Agilent-Varian Mercury 400-MHz spectrometer with DMSO-d6 as the solvent. All coupling constants are measured in hertz, and the chemical shifts (δH) are quoted in parts per million relative to TMS (δ0), which was used as the internal standard. High-resolution MS was carried out on an Agilent 6210 LC-MS (electrospray ionization-TOF) system. HPLC analysis was performed using a JASCO HPLC system equipped with a PU-2089 Plus quaternary gradient pump and a UV-2075 Plus UV-visible detector, using an Alltech Kromasil C-18 column (150 × 4.6 mm, 5 μM) and an Agilent Eclipse XDB-C18 column (150 × 4.6 mm, 5 μM). The purities of the final compounds used for the biochemical and functional studies were >95% as measured by HPLC. Melting points were recorded on an Optimelt automated melting point system (Stanford Research Systems). TLC was performed using silica gel 60 F254 plates (Thermo Fisher Scientific), with observation under UV when necessary. Anhydrous dimethylformamide was used as purchased from Sigma-Aldrich. Burdick and Jackson HPLC-grade solvents were purchased from VWR for HPLC, HPLC-MS, and high-resolution mass analysis. dBET1 (HPLC purity 98%) was prepared from JQ1 as described(296). Detailed information about the synthesis of N-methyl-Len, and N-methyl-dBET1 synthesis are provided in the supporting material.
3.4.6 Cloning, protein expression, and purification

The full-length hCRBN protein (isoform 1) in complex with DDB1 was a generous gift from Celgene Corp. (San Diego, CA). The gene coding for the human TBD (amino acids 319–425) was synthesized and subcloned into the BamHI-NotI restriction sites of the pGEX-6P-1 vector by GeneArt® gene synthesis. The gene was engineered with silent mutations that utilize the favored E. coli codons. TBD E377V, V388I, H378A, and W380A mutations were performed using PCR. Details of primer sequences are provided in the supporting material. Mutations were confirmed by sequencing. The recombinant DNA plasmids were transformed into E. coli Rosetta™ 2(DE3)pLysS competent cells (EMD Millipore, Billerica, MA) for subsequent protein expression. The GST-tagged TBD proteins linked with PreScission protease proteolytic site were expressed and purified as follows. A single colony of freshly transformed cells was cultured at 37 °C for 16 h in 5 ml of Luria–Bertani (Thermo Fisher Scientific) medium containing 100 μg/ml ampicillin (Sigma-Aldrich) and 34 μg/ml chloramphenicol (Sigma-Aldrich). 1 ml of the culture was then used to inoculate 25 ml of Terrific Broth–phosphate medium (Thermo Fisher Scientific) with 100 μg/ml ampicillin at 37 °C for 16 h. The culture was then transferred to 1.5 liters of Terrific Broth–phosphate medium supplemented with 50 μM ZnCl2 (Sigma-Aldrich). The resultant culture was incubated with continuous shaking at 250 rpm to an A600 of 0.70 and then induced with isopropyl-β-d-thiogalactopyranoside (0.5 mm final concentration; Thermo Fisher Scientific) at 16 °C for 20 h before harvesting by centrifugation at 6000 rpm for 30 min. The cells were lysed by homogenization in 50 mm Tris (pH 8.0; Sigma-Aldrich), 500 mm NaCl (Thermo Fisher Scientific), 1 mm TCEP (Sigma-Aldrich), 0.1% Triton X-100 (Sigma-Aldrich), 10 μM ZnCl2 (Acros Organics, Thermo Fisher Scientific), and protease inhibitor mixture (Roche Applied Science). The protein was then purified by affinity chromatography on an AKTA
Explorer or AKTA Purifier (GE Healthcare Life Sciences) using a GSH-Sepharose matrix (GE Healthcare) pre-equilibrated with 50 mm Tris (pH 8.0), 500 mm NaCl, 1 mm TCEP, and 10 μM ZnCl2 and eluted with the same buffer with the addition of 10 mm reduced GSH (Sigma-Aldrich). Purity of the protein in the different fractions was determined by SDS-PAGE, and the best fractions were pooled. GST was cleaved from the pooled GST-TBD fractions by digestion with PreScission protease at 4 °C for 4 h. GST was removed from the resultant digest by a second round of GST affinity chromatography. Proteins were further purified by size-exclusion chromatography in a Superdex 75 column (GE Healthcare Life Sciences). Fractions with >90% purity were pooled, concentrated by ultrafiltration (10K Amicon tubes, EMD Millipore), and stored at −80 °C.

### 3.4.7 Zincon assay

All chemicals used in this assay were purchased from Sigma-Aldrich. The assay was adapted from previously published methods (309). A zinc concentration standard curve was prepared in 50 mm borate buffer, pH 9.0, containing 4 m NaCl, 8 m urea, and 40 μM zincon (2-carboxy-2′-hydroxy-5′-sulfoformazylbenzene) dye. Purified proteins were acidified with 300 mm HCl to facilitate the release of the zinc ions bound to the protein. The protein polypeptide was separated from the water-soluble layer by centrifugation. The solution was then spiked with 10–20 μM zinc sulfate. Absorption spectra between 400 and 750 nm were recorded. λmax of free zincon and zincon–zinc complex were measured at 480 and 620–630 nm, respectively. Absorbance at 630 nm of different zinc concentrations was used to generate the linear regression curve. The concentration of zinc-containing protein was extrapolated based on the linear regression curve.
3.4.8 ITC

The binding of CRBN–DDB1 complex and CRBN–TBD WT and mutant variants to immunomodulatory compounds was analyzed with a MicroCal iTC200 titration calorimeter (Malvern, Westborough, MA). The compound phthalimide was used as the negative control. The proteins were rebuffered into binding buffer (50 mm HEPES (pH 7.5, Sigma-Aldrich), 200 mm NaCl, 0.1 mm TCEP, and 0.6% DMSO). For the titrations of the protein constructs, a total of 19 aliquots (2.05 μl each) of the respective compounds (~600 μM) were injected into 200 μl of the protein solutions (40 μM) at 25 °C. The ITC cell mixture was constantly stirred at 1000 rpm and recorded for 160 s between injections at low feedback. The corrected heat values were fitted using a nonlinear least square curve-fitting algorithm (Microcal Origin version 7.0, OriginLab, Northampton, MA) to obtain binding constants (K_D) and values for n (number of binding sites), ΔH (enthalpy), and ΔS (entropy).

3.4.9 Intrinsic tryptophan fluorescence assay

Binding of compounds to WT and mutant TBD was monitored by fluorescence spectroscopy, using an adapted previously published method (288,317). All chemicals used in this assay were purchased from Sigma-Aldrich unless otherwise stated. In this assay, changes in emission spectra are induced by interactions of these compounds with the three Trp residues (Trp-380, Trp-386, and Trp-400) in the binding site (307,308). TBD proteins (final concentration, 10 μM) were incubated with varying final concentrations (0–750 μM) of compounds in assay buffer (50 mm Tris, pH 7.5, 200 mm NaCl, 0.1% Pluronic-F127, and 1 mm TCEP) to a final volume of 40 μl in a black 96-well half-area plate (Corning, Inc.). A 0.5% final DMSO concentration was used in each well. Samples were excited at 280 nm, and fluorescence emission
intensities were measured at 340 nm using a Wallac Envision 2102 multilabel plate reader (PerkinElmer Life Sciences). All measurements were done in triplicate and corrected for inner filter effect to subtract for ligand-associated fluorescence, as described (326). The magnitude of fluorescence difference \((1 - F/F_0)\) was measured, where \(F\) is the fluorescence emission at a given concentration of ligand; \(F_0\) is the intrinsic fluorescence intensity of 10 \(\mu\)M TBD protein alone. Graph plotting and curve fitting to obtain apparent dissociation constant (\(K_D\)) values were calculated by fitting the relative change in intrinsic fluorescence at 340 nm \((1 - F/F_0)\) versus ligand concentration to a nonlinear regression with one-site binding hyperbola with GraphPad Prism (GraphPad Software, La Jolla, CA).

### 3.4.10 Molecular Modeling Preparation

The preparation of the protein systems for hCRBN (PDB code 4TZ4 (307), mCRBN (PDB codes 4TZC and 4TZU) (307), and gCRBN (PDB codes 4CI1, 4CI2, and 4CI3) (308) were done using the Schrödinger software suite (Maestro, version 9.7, Schrödinger, LLC, New York). Protein structure coordinates were downloaded from the PDB (227,228) and prepared with the Protein Preparation Wizard (PrepWizard) in Maestro (Schrödinger Suite 2014-1 Protein Preparation Wizard; Epik version 2.7, Schrödinger; Impact version 6.2, Schrödinger; Prime version 3.5, Schrödinger) (231). Final system equilibration was determined by the observation of asymptotic behavior of the potential energy, RMSD, and Rg profiles and visual inspection of trajectories guided by root mean square fluctuation profiles (See Section 3.8.1 Supplementary Computational Results). MD simulations were performed with the Desmond MD program with additional details provided in the supporting Materials and Methods (Desmond Molecular Dynamics System, version 3). After equilibration was determined, a hierarchical average linkage
clustering method based on RMSD was utilized to determine an average representative structure for each equilibrated system. The program PROPKA was then implemented again on the equilibrated structures to test the consistency of side chain protonation states at pH 7.4 (Section 3.8.3 Determination of H378 Protonation States provides details for characterization of the ionization states for certain side chains).

3.5 Author Contributions

A. A. A., A. Y. A., and R. M. contributed equally to the manuscript and designed the research, performed experiments, analyzed and interpreted data, prepared figures, and wrote the manuscript. M. S. B. performed experiments. A. B. designed the research plan, performed experiments, analyzed and interpreted data, and prepared the paper. J. M. M. performed experiments and analyzed and interpreted data. R. S. H. performed experiments, analyzed and interpreted data, and reviewed the paper. W. E. G. performed experiments, analyzed data, and prepared figures. S. G. designed the research, performed experiments, analyzed and interpreted data, and prepared figures. M. A. synthesized chemicals. Y. Y. produced protein for the project. M. R. K. performed experiments. M. E. O. conducted experiments. K. D. provided concept development for modeling, established experimental methods, and prepared the paper. W. G. provided concept development and provided expertise for modeling. J. A. Y. provided concept development and prepared the manuscript. A. M. R. provided critical reagents. E. S. provided assistance with experimental design, interpreted data, and prepared the paper. H. R. L. provided concept design for key reagents, interpreted results, and prepared the paper. N. J. L. and P. K. E.-B. shared equally in manuscript preparation and designed the research plan, performed data analysis.
analysis and interpreted the results, prepared the paper, and were responsible for all aspects of
the work.

3.6 Acknowledgments

hCRBN–DDB1 protein complex and the anti-CRBN antibody were generous gifts from
Celgene Corp. (San Diego, CA). Multiple myeloma cell lines were provided by Drs. Kenneth
Shain and Connor Lynch (Moffitt Cancer Center). We also thank John Cleveland, Ph.D. (Moffitt
Cancer Center) for his expert opinion and assistance with manuscript preparation. The Chemical
Biology Core Facility at the H. Lee Moffitt Cancer Center and Research Institute, is an NCI
(National Institutes of Health)-designated Comprehensive Cancer Center (Grant P30-
CA076292).

This work was supported through a Research Agreement with Celgene Corp. J. M. M. is
currently a salaried employee of Celgene Corporation but was a graduate student in the Cancer
Biology Ph.D. program at the time of this study. Dr. N. Lawrence and the Moffitt Cancer Center
have received payment from licensing unrelated technology to Celgene and have appropriate
conflict-of-interest management plans.

The abbreviations used are:

CRBN cereblon

hCRBN human CRBN
mCRBN mouse CRBN
gCRBN chicken CRBN
Len lenalidomide
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pom</td>
<td>pomalidomide</td>
</tr>
<tr>
<td>MDS</td>
<td>myelodysplastic syndrome</td>
</tr>
<tr>
<td>TBD</td>
<td>thalidomide-binding domain</td>
</tr>
<tr>
<td>BET</td>
<td>bromodomain and extraterminal domain</td>
</tr>
<tr>
<td>PROTAC</td>
<td>proteolysis targeting chimera</td>
</tr>
<tr>
<td>TCR</td>
<td>T-cell receptor</td>
</tr>
<tr>
<td>RMSD</td>
<td>root mean square deviation</td>
</tr>
<tr>
<td>MD</td>
<td>molecular dynamics</td>
</tr>
<tr>
<td>DDB1</td>
<td>DNA damage–binding protein 1</td>
</tr>
<tr>
<td>IFD</td>
<td>induced-fit docking</td>
</tr>
<tr>
<td>IF</td>
<td>intrinsic tryptophan fluorescence assay</td>
</tr>
<tr>
<td>ITC</td>
<td>isothermal titration calorimetry</td>
</tr>
<tr>
<td>CTV</td>
<td>cell trace violet</td>
</tr>
<tr>
<td>TCEP</td>
<td>tris(2-carboxyethyl)phosphine.</td>
</tr>
</tbody>
</table>

3.7 Supplementary Material

3.7.1 Supplementary Methods

3.7.1.1 Protein Preparation. The preparation of the protein systems for human cereblon (hCRBN, PDB 4TZ4 (307)), murine cereblon (mCRBN, PDB 4TZC and 4TZU (307)), and galline cereblon (gCRBN, PDB 4CI1, 4CI2, and 4CI3 (308)) were done using the Schrodinger software suite (25). Protein structure coordinates were downloaded from the Protein Data Bank.
and prepared with the Protein Preparation Wizard (PrepWizard) in Maestro. Bond orders were assigned, including disulfide bridges, and original hydrogens were deleted and later replaced to reduce bad contacts and other crystal artifacts before protonation and hydrogen bond optimization. Missing side chains were added and optimized using Prime (16,66,67). Cofactors used in crystallization (such as sulfate or phosphate ions), ligands, and additional protein dimers were then deleted. All waters were retained for assisting in the determination of side chain protonation states and initial hydrogen bond optimization. Hydrogen atoms were then added to the protein, remaining cofactors, and to any added structural waters. The program PROPKA (50) was used for the prediction of protein ionization states at 7.4 pH and ProtAssign was used for hydrogen bond optimization. After automatic hydrogen assignment, visual inspection was used to flip residues and change protonation states at protein-protein interfaces when appropriate. Specific attention was given to the tautomeric states for His residues, which were assumed to be neutral, and potential metal-igation and hydrogen-bonding interactions were considered. The Nδ nitrogen for H378 was kept protonated for each system. Four of the six crystallographic structures of CRBN used in this study appear to show the H378 δ-nitogen coordinated with the carbonyl oxygen of the ligand. This posits the assumption of a hydrogen bond and that the H378 δ-nitrogen must therefore be protonated (327–331). PROPKA analysis before and after simulations also confirm this assumption. As such, all MD systems were generated with the neutral H378 δ-nitrogen as being protonated. All waters more than 3 Å from heterogen (HET) groups were then removed.

3.7.1.2 Molecular Dynamics. hCRBN and gCRBN were co-crystallized with DNA damage-binding protein 1 (DDB1), whereas mCRBN was truncated (108 defined residues for
mCRBN compared to 380 defined residues for hCRBN and gCRBN) and monomeric (i.e. not complexed with DDB1). DDB1 was included in the simulations for hCRBN and gCRBN. The thalidomide binding domain for hCRBN and mCRBN differs by only three residues. Therefore, to further increase sampling and determine any structural dependence on these residues, the original hCRBN sequence of the equilibrated representative structure was mutated to the mCRBN sequence with the following mutations, C366S, E377V, and V388I to form a hCRBN to mCRBN mutant (hmCRBN). Another system was also constructed from the hCRBN representative structure with the Nε nitrogen on the H378 residue protonated (hCRBN-pNε) to compare and provide insight into the protonation states of H378. After protein refinement, a total of five protein systems (hCRBN, hCRBN-pNε, mCRBN, gCRBN, and hmCRBN) were constructed for MD simulation.

MD simulations were performed with the Desmond MD program (233–236). Periodic boundary conditions were imposed on a cubic simulation box extending at least 10Å from the protein. The simulation box was then solvated with TIP3P waters (237), electrically neutralized by introducing sodium ions, and the OPLS-2005 all-atom force field (332,333) was applied to all atoms. The SHAKE algorithm (238) was used to constrain all bonds in the system and the REference System Propagator Algorithm (RESPA) (239) with an integration time step of 2 fs was employed. Long-range electrostatics were calculated using the Particle Mesh Ewald (PME) algorithm(88,89) with a real-space cutoff of 13 Å. A cutoff of 16 Å was employed for van der Waals interactions. The systems were simulated in an NPT ensemble using the Martyna-Tuckerman-Tobias-Klein (MTTK) barostat(241,242) at a constant pressure of 1 atm and a temperature of 310 K using the Nose–Hoover temperature coupling scheme(240).
All systems were energy minimized with a truncated newtonian conjugate gradient (TNCG) method (243) followed by multiple restrained minimizations to randomize systems before equilibration and final simulation. Heavy atoms of the protein were held fixed during heating for an initial 12 ps NVT ensemble simulation at 10 K with the Berendsen thermostat (244). This was followed by simulations at 1 atm in the NPT ensemble for 12 ps at 10 K and 24 ps at 310 K. Unrestrained equilibration MD was then performed for 24 ps at 310 K and 1 atm. Finally, unconstrained production MD was performed on the hCRBN and mCRBN systems for 100 ns. Due to the stability observed in hCRBN, its close similarity to gCRBN, and since the hCRBN-pNε system was derived from the equilibrated hCRBN system, the gCRBN and hCRBN-pNε systems were only run for 50 ns. Energies were recorded every 2 ps and trajectory frames were recorded every 5 ps.

Final system equilibration was determined by the observation of asymptotic behavior of the potential energy, Root Mean Square Deviation (RMSD), and Radius of Gyration (Rg) profiles and visual inspection of trajectories guided by Root Mean Square Fluctuation (RMSF) profiles (Supplementary Material).

3.7.1.3 Consensus Docking. After equilibration was determined, a hierarchical average linkage clustering method based on RMSD was utilized to determine an average representative structure for each equilibrated system. The program PROPKA was then implemented again on the equilibrated structures to test consistency of side chain protonation states at 7.4 pH. The representative structure was then used for consensus docking incorporating five diverse and complimentary docking methods described below. By applying these varied energy scoring methods, the weaknesses of each method can be identified for a particular model and error
statistically minimized, yielding a more accurate summary of ligand binding dispositions and affinities.

As a check for the placement of the grid used in the docking studies and for further analysis of the binding cavity, Schrödinger's SiteMap program (245–247) was employed. SiteMap searches the protein structure for likely binding sites and highlights regions within the binding site suitable for occupancy by hydrophobic groups, hydrogen-bond donors, acceptors, or metal-binding functionality of the ligand.

The ligands thalidomide, pomalidomide, and lenalidomide were prepared using the program LigPrep (248) and the OPLS-2005 all-atom force field was applied to all ligand atoms.

### 3.7.1.4 Induced Fit Docking

The representative structures were then used for docking with the induced-fit docking (IFD) method in the Schrödinger software suite (122–124,334). The IFD methodology incorporates both the docking program Glide (114,121,125,335) to account for ligand flexibility and the Refinement module in the Prime program to account for receptor flexibility. The Schrödinger IFD protocol attempts to model induced-fit effects from alterations in binding site conformation due to ligand binding in order to increase accuracy of binding affinity estimates and prediction of possible binding modes.

The position of the cubic docking grid was centered on the original co-crystallized ligand centroids and was given a size of 29 Å. A constrained minimization of the receptor was performed with an RMSD cutoff of 0.18 Å. An initial softened potential Glide docking of the ligand set was then implemented with the standard precision (SP) mode and a van der Waals scaling factor of 0.5 was applied to the non-polar atoms of the receptor and ligands. The resulting top 20 poses of the ligands were used to sample protein plasticity by conformational searches and minimizations of binding pocket residues within 6 Å of any ligand pose for all complexes.
obtained. The new receptor conformations were then redocked using complexes within 30 kcal/mol from the best scoring structure. Glide docking parameters for this step were reset to the default hard potential function with a van der Waals scaling of 1.0 and SP mode.

The estimated binding affinity of each complex was reported in the GlideScore and used to compare differences between each ligand while the Emodel score is used to inter-compare poses of the ligands. Emodel places more significance on weighting force field components (electrostatic and van der Waals energies), making it better for comparing conformers as opposed to comparing chemically-distinct species.

### 3.7.1.5 Rigid Receptor Docking (RRD)

Rigid docking simulations were performed by Glide (114,121,125). Glide uses a GlideScore fitness function based on Chemscore (249,250) for estimating binding affinity, but includes a steric-clash term, adds buried polar terms to penalize electrostatic mismatches, and modifies other secondary terms. Docking simulations used both the standard precision (SP) and extra precision (XP) methods. XP mode is a refinement algorithm enforced only on good ligand poses. Sampling is based on an anchor and refined growth strategy and the scoring function includes a more complete treatment of some of the SP energetic terms, such as the solvation and hydrophobic terms. Docking grids were defined by a rectangular ligand atom inclusion outer box of 22Å and ligand centroid constraint inner box of 10Å in the x, y, and z directions originating from the binding cavity centroid defined by the original co-crystallized IMiD ligand centroid for the CRBN TBD.

### 3.7.1.6 Quantum Polarized Ligand Docking (QPLD)

To account for ligand polarization upon binding, Quantum Mechanics/Molecular Mechanics (QM/MM) docking was performed by the Schrödinger QM-Polarized Ligand Docking Protocol (QPLD) (126,127). The
protocol first employs RRD using Glide in SP mode. In this step, the top five poses of each ligand in the initial RRD were used. Potential ligand polarization induced by the protein were then calculated with Qsite (111,251,252) at the B3LYP/6-31G* level. The ligand force fields were then reconstructed with QM/MM modified charges, redocked, and five poses of each ligand were saved for evaluation.

3.7.1.7 Molecular Mechanics and Generalized Born Surface Area (MM/GBSA). The MM/GBSA method combines molecular mechanics energy terms and implicit solvation models to calculate the binding-free energy based on docking complexes. The protocol, implemented by the Prime MM-GBSA module, calculates optimized free energies for the free protein and free ligand and references them with the original bound complex energy (253). Polar contributions are calculated using the Generalized Born (GB) model (254), an implicit solvent model is based on a variable dielectric surface Generalized Born (VD-SGB) approach, where the variable dielectric value for each residue was fit to a large number of side-chain and loop predictions while the non-polar energy is estimated using the solvent accessible surface area (SASA) (255). The simulation was performed based on receptor–ligand complex structures obtained from induced fit docking. The obtained ligand poses were minimized using the local optimization feature in Prime, whereas the energies of complex were calculated with the OPLS-3 force field and Generalized-Born/Surface Area continuum solvent model (256). During the simulation process, the ligand strain energy is also considered. A known issue with MM/GBSA is that scores do not accurately reproduce absolute physical binding affinities but display great efficacy at ranking compounds in a relative manner (257–260). We developed a correlation function using a single-layer logistic regression to rescale MM/GBSA scores based on the other docking score
algorithms. This retains the ranking accuracy of MM/GBSA and allows us to proportion the results in a minimally biased and physically relevant manner.

Figure 31. Synthesis of tert-Butyl (2-(2,6-dioxopiperidin-3-yl)-1-oxoisoindolin-4-yl)carbamate.

3.7.1.8 Detailed Chemical Methods. Lenalidomide (259 mg, 1 mmol) and Boc₂O (218 mg, 1.1 mmol) were mixed in THF (1 mL) in a sealed tube and stirred at 60 °C overnight. The next day, Boc₂O (110 mg, 0.5 equiv.), THF (1 mL), and DMF (0.5 mL) were added and the solution was further stirred at 120 °C overnight. Water (20 mL) was added and the mixture was sonicated. The precipitate was filtered, washed with water (10 mL), and dried. The resulting solid was triturated using EtOH/EtOAc/hexanes and filtered to give the desired product as an off-white solid (258 mg, 72%). Mp: 196–198 °C. HPLC–MS (ESI+): m/z 741.3 [(100%, 2M+Na) + ], 719.4 [(40%, 2M+H)+ ], 360.2 [(90%, M+H)+ ]. 1 H NMR (400 MHz, DMSO-d6): δ 11.00 (s, 1H), 9.21 (s, 1H), 7.74 (dd, J = 6.8, 1.7 Hz, 1H), 7.49–7.39 (m, 2H), 5.10 (dd, J = 13.3, 4.7
Hz, 1H), 4.41 (d, J = 17.6 Hz, 1H), 4.32 (d, J = 17.6 Hz, 1H), 2.95–2.83 (m, 1H), 2.64–2.54 (m, 1H), 2.40–2.26 (m, 1H), 2.05–1.95 (m, 1H), 1.46 (s, 9H). Compound 1 was reported previously (296). tert-Butyl (2-(1-methyl-2,6-dioxopiperidin-3-yl)-1-oxoisooindolin-4-yl)carbamate (2): To a mixture of 1 (100 mg, 0.278 mmol) and K2CO3 (38 mg, 0.278 mmol) in DMF (0.8 mL) was added methyl iodide (0.017 mL, 0.278 mmol) dropwise at room temperature under Argon. The mixture was stirred overnight. Water (10 mL) was added and extracted with EtOAc (2 × 20 mL). The combined organic layers were dried (Na2SO4), filtered, and concentrated under reduced pressure. The resulting yellow oil was purified by flash chromatography (SiO2) eluting with hexanes in EtOAc (80% to 100%) 4 M HCl in dioxane r.t., 3.5 h 75% 3 N O NH2 NH O O N O NHBoc NH O O Boc2O, THF/DMF (4:1) sealed tube 60-120 °C, 2 d N O NHBoc N O O MeI, K2CO3 DMF, r.t., o/n N O NH2 N O O Lenalidomide 72% 1 39% 2 HCl to provide the title compound as a white solid (40.37 mg, 39%). Mp: 192 °C (dec). HPLC–MS (ESI+): m/z 741.3 [(100%, 2M+Na)+ ], 719.4 [(40%, 2M+H)+ ], 360.2 [(90%, M+H)+ ]. 1 H NMR (400 MHz, DMSO-d6): δ 9.20 (s, 1H), 7.73 (dd, J = 6.5, 2.3 Hz, 1H), 7.49–7.41 (m, 2H), 5.16 (dd, J = 13.4, 5.1 Hz, 1H), 4.42 (d, J = 17.6 Hz, 1H), 4.30 (d, J = 17.6 Hz, 1H), 3.02–2.90 (m, 1H), 2.99 (s, 3H), 2.80–2.71 (m, 1H), 2.40–2.27 (m, 1H), 2.07–1.97 (m, 1H), 1.46 (s, 9H). 3-(4-Amino-1-oxoisooindolin-2-yl)-1-methylpiperidine-2,6-dione (3 or N1 -methyl-lenalidomide): 2 (35 mg, 0.093 mmol) was stirred in 4 M HCl in dioxane (0.5 mL) for 3.5 h at room temperature. The white suspension was concentrated under reduced pressure and the resulting solid was triturated in DCM/hexanes, washed with EtOAc and hexanes (10 mL each), and dried to provide the title compound as light yellow flakes (21.81 mg, 75%). Mp: 207 °C (dec). HPLC: 99% [tR = 11.6 min, 10% MeOH, 90% water (with 0.1% TFA), 20 min]. 1 H NMR (400 MHz, DMSO-d6): δ 7.29 (t, J = 7.5 Hz, 1H), 7.11 (d, J = 7.5 Hz, 1H), 6.99 (d, J = 7.5 Hz, 1H), 5.17 (dd, J = 13.4, 4.7
Hz, 1H), 5.20–4.80 (br s, 2H, disappeared on D2O shake), 4.28 (d, J = 17.0 Hz, 1H), 4.17 (d, J = 17.0 Hz, 1H), 3.05–2.91 (m, 1H), 2.99 (s, 3H), 2.80–2.70 (m, 1H), 2.39–2.25 (m, 1H), 2.08–1.97 (m, 1H). HPLC–MS (ESI+): m/z 569.2 [(30%, 2M+Na)+ ], 274.2 [(100%, M+H)+ ]. LC–MS (ESI+): 569.2 [40%, (2M+Na)+ ], 296.1 [100%, (M+Na)+ ]. HRMS (ESI+): m/z calcd for C14H15N3O3 (M+H)+ 274.1186, found 274.1176.

Figure 32. Structure of Me-dBET1.

2-((S)-4-(4-Chlorophenyl)-2,3,9-trimethyl-6H-thieno[3,2-f][1,2,4]triazolo[4,3-a][1,4]diazepin-6-yl)-N-(4-(2-(1-methyl-2,6-dioxopiperidin-3-yl)-1,3-dioxoisindolin-4-yl)oxy)acetamido)-butyl)acetamide (Me-dBET1): This was prepared from N-Boc aminoglutarimide (details to be reported elsewhere) as a white solid. HPLC: 99% [tR = 24.7 min, gradient 5-95%, MeOH-water (with 0.1% TFA), 30 min]. 1H NMR (400 MHz, Methanol-d4) δ 7.82 (dd, J = 8.4, 7.4, Hz, 1H), 7.53 (dd, J = 7.2, 0.4 Hz, 1H), 7.47–7.38 (m, 5H), 5.14 (dt, J = 12.9, 5.1 Hz, 1H), 4.78 (s, 2H), 4.62 (ddd, J = 9.0, 5.2, 0.7 Hz, 1H), 3.44–3.34 (m, 3H), 3.30–3.22 (m, 3H), 3.11 (2 xs, 3H), 2.86–2.82 (m, 2H), 2.73–2.62 (m, 4H), 2.43 (2 x d J = 0.9 Hz, 3H), 2.12–2.03 (m, 1H), 1.72–
1.69 (m, 3H), 1.67–1.59 (m, 4H); HPLC–MS (ESI+): m/z 821.2 [(100%, M+Na)+], 799.3 [(70%, M+H)+]; HRMS (ESI+): m/z calcd for C39H39ClN8O7S (M)+ 798.2345, found 798.2368.

3.7.1.9 Detailed Protein Production of Wild-Type and Mutant CRBN. The primers used are

5'-CCAGTCTGTGTGTTGAAACAGGCAAGAAACC-3' and 5'-GGTTTCTTTGCTCTGTTTACAA CACAGACTGG-3';

5'-GGTTATGCATGACCACACGACAGTGTTAATTTGTGC-3' and 5'-GCACAAAATTTTACACTGTGCATGGTGCCATGCATAAC-3',

5'-CGTCCGAGCACCAGAAGCAAG CTGGTTTCCGGGTATGC-3' and 5'-GCATAAACCAGAAACCAGGTGCTCGGTGCTCG-3';

5'-CGAGCACCAGAACCAGCTGTTCCGGGTATGCATGG-3' and 5'-CCATGCATAA CCCGAAACCGCCTATGTTCGTCCTCG-3', respectively.
Figure 33. Sequence conservation of CRBN thalidomide binding site. (A) Diagrammatic representation of the CRBN’s ubiquitin targeting complex containing DDB1, Cul4A, Roc1, Rbx1. The region of the thalidomide-binding domain (TBD) indicated (red) interacts with the glutarimide ring of the IMiDs, whilst the second ring (phthalimide, isoindolinone and quinazolinone) interacts with the CRBN TBD and its substrates (Ikaros, Aiolos, CK1α, and unknown proteins), respectively. Also shown is the structure of dBET1 that recruits BRD4 and other BET-domain containing proteins and molecules that interact with the JQ1-like BET targeting group. (B) The CRBN protein sequences from diverse vertebrate species that correspond to the human thalidomide binding site were aligned using Clustal Omega (Sievers
and Higgins 2014). Positions that are at least 90% identical are shaded in black and similar residues are shaded in gray. Shading in other colors highlight sequence diversity at three positions: Cys366, Glu377 and Val388 in human (above) and Ser369, Val380 and Ile391 in mouse (below). Arrows indicate cysteines involved in forming disulfide bridges. Sequences include: thirteen primates (human, NP_001166953.1, common chimpanzee, XP_001140433.1; western lowland gorilla, XP_004033566.1; Sumatran orangutan, NP_001127555.1; northern white-cheeked gibbon, XP_012357648.1; rhesus macaque, NP_001182576.1; olive baboon, XP_003894216.1; green monkey, XP_007983240.1; squirrel monkey, XP_010336170.1; common marmoset, XP_008980298.1; Philippine tarsier, XP_021562663.1; gray mouse lemur, XP_020139859.1; Sunda flying lemur, XP_008581608.1), five non-primate mammals (mouse, NP_067424.2; rat, NP_001015003.1; domestic dog, XP_005632293.1; cow, NP_001068995.1; Florida manatee, XP_012409572.1), one marsupial (gray short-tailed opossum, XP_001374178.2), one reptile (American alligator, XP_006263115.1), two birds (chicken, XP_004944767.1; zebra finch, XP_012429169.1), one amphibian (Xenopus, NP_001008192.1), one lobe-finned fish (West Indian Ocean coelacanth, XP_006001868.1), two rayfinned fish (Spotted gar, XP_006630756.1; zebrafish, NP_001003996.1) and one cartilaginous fish (whale shark, XP_020385559.1).
Figure 34. Dose of anti-CD3 and lenalidomide to induce IL-2 and proliferation by T cells. (A) Relative IL-2 mRNA production with increasing concentrations anti-CD3 (0.01 – 10 μg/mL) stimulation. (B) IL-2 secretion by ELISA in unstimulated cells and after treatment with anti-CD3 (5μg/mL) and anti-CD28 (1 μg/ml) with increasing concentrations of lenalidomide. (C) Proliferation in human T cells and (D) mouse T cells treated with DMSO (vehicle) or lenalidomide (10 μM) measured by S-phase transition as indicated by incorporation of bromodeoxyuridine (BrdU) with detection by flow-cytometry in cells stained for CD8+ surface expression.
Table 11. IC$_{50}$ (μM) of Multiple Myeloma Cell Lines

<table>
<thead>
<tr>
<th>Drug</th>
<th>H929</th>
<th>U266</th>
<th>MM1.S</th>
<th>5TGM1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lenalidomide (Len)</td>
<td>5.13</td>
<td>10.0</td>
<td>1.48</td>
<td>&gt;&gt;100</td>
</tr>
<tr>
<td>Pomalidomide (Pom)</td>
<td>0.046</td>
<td>0.51</td>
<td>0.099</td>
<td>&gt;&gt;100</td>
</tr>
</tbody>
</table>

H929, U266 and MM1.S are human multiple myeloma cell lines while 5TGM1 is a mouse multiple myeloma cell line. IC$_{50}$ (μM) calculated as follows: log(inhibitor) vs response based on variable slope (four parameters), Y=bottom (best fit value) + (100-Bottom)/1+10^((LogEC50-X)*n)).

Figure 35. RMSD profiles of hCRBN, mCRBN, gCRBN, and hmCRBN. (A) RMSD profiles of hCRBN. Profiles were created using backbone atoms and were calculated with hCRBN-DDB1 complex (Bottom Graph), hCRBN alone (Middle Graph), and binding site residues N335 to A421 (Top Graph). (B) RMSD profiles of mCRBN. Profiles were created using backbone atoms and were calculated with all hCRBN residues (Bottom Graph), binding site residues N335 to A421 (Middle Graph), and residues (357, 377, 379-383, 388-390, 401, 402, 404) located 6 Å away from the ligand (Top Graph). (C) RMSD profiles of gCRBN. Profiles were created using backbone atoms and were calculated with hCRBN-DDB1 complex (Bottom Graph), gCRBN alone (Middle Graph), and binding site residues N337 to A423 (Top Graph). (D) RMSD profiles of hmCRBN. Profiles were created using backbone atoms and were calculated with hmCRBN-DDB1 complex (Bottom Graph), hmCRBN alone (Middle Graph), and binding site residues N335 to A421 (Top Graph). (E) Rg Profiles of hCRBN. Profiles were created from all atoms and were calculated with hCRBN-DDB1 complex (Bottom Graph) and hCRBN alone (Middle Graph). (F) Rg Profiles of mCRBN. Profiles were created from all atoms. (G) Rg Profiles of gCRBN. Profiles were created from all atoms and were calculated with gCRBN-DDB1 complex (Bottom Graph) and hCRBN alone (Top Graph). (H) Rg Profiles of hmCRBN. Profiles were created from all atoms and were calculated with hmCRBN-DDB1 complex (Bottom Graph) and hmCRBN alone (Top Graph). All measurement units are in Angstroms.
Figure 36. RMSF profiles of hCRBN, mCRBN, gCRBN, and hmCRBN. (I) RMSF profile of hCRBN. Profiles were created using backbone atoms and were calculated with hCRBN alone (DDB1 RMSF available upon request). Binding site residues N335 to A421 displayed as dashed line. (J) RMSF Profile of mCRBN. Profiles were created using backbone atoms of mCRBN. Binding site residues N335 to A421 displayed as dashed line. (K) RMSF Profile of gCRBN. Profiles were created using backbone atoms and were calculated with gCRBN alone (DDB1 RMSF available upon request). Binding site residues N337 to A423 displayed as dashed line. (L) RMSF Profile of hmCRBN. Profiles were created using backbone atoms and were calculated with hmCRBN alone (DDB1 RMSF available upon request). Binding site residues N335 to A421 displayed as dashed line. All measurement units are in Angstroms.

Figure 37. Potential energy profiles of hCRBN, mCRBN, gCRBN, and hmCRBN (M) Potential energy profile of hCRBN simulation, (N) mCRBN simulation, (O) gCRBN simulation, and (P) hmCRBN simulation. Multiple linear fits were characterized to determine asymptotic behavior by minimizing slope.
Figure 38. Hydrogen bonding analysis of Len isoindolinone amine and E377 carboxyl group. Top two graphs show inter-atomic distance between ligand nitrogen and glutamate side chain oxygens with line denoting maximum hydrogen bond cutoff of 3.5 Å. Bottom graph indicates programmatic detection of hydrogen bonding incorporating maximum NHO angle cutoff of 90°.
Table 12. Pose RMSD Calculations and Docking Scores.

<table>
<thead>
<tr>
<th></th>
<th>hmCRBN Post MD</th>
<th>hCRBN Post MD</th>
<th>hCRBN Crystal</th>
<th>gCRBN Post MD</th>
<th>gCRBN Crystal</th>
<th>mCRBN Post MD</th>
<th>mCRBN Crystal</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Thalidomide</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Emodel</td>
<td>-81.796</td>
<td>-83.437</td>
<td>-78.630</td>
<td>-77.239</td>
<td>-79.853</td>
<td>-63.197</td>
<td>-63.881</td>
</tr>
<tr>
<td>Pose RMSD</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.6615\textsuperscript{a}</td>
<td>1.1036\textsuperscript{a}</td>
<td>7.1958\textsuperscript{b}</td>
<td>7.0379\textsuperscript{b}</td>
</tr>
<tr>
<td><strong>Lenalidomide</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Emodel</td>
<td>-85.454</td>
<td>-87.014</td>
<td>-79.857</td>
<td>-78.197</td>
<td>-77.976</td>
<td>-52.240</td>
<td>-57.118</td>
</tr>
<tr>
<td>Pose RMSD</td>
<td>0.5944\textsuperscript{c}</td>
<td>0.7384\textsuperscript{c}</td>
<td>1.7233\textsuperscript{c}</td>
<td>1.5859\textsuperscript{d}</td>
<td>1.0053\textsuperscript{d}</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>Pomalidomide</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Emodel</td>
<td>-89.915</td>
<td>-86.330</td>
<td>-81.008</td>
<td>-80.357</td>
<td>-84.483</td>
<td>-51.789</td>
<td>-58.872</td>
</tr>
<tr>
<td>Pose RMSD</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.7592\textsuperscript{e}</td>
<td>0.9482\textsuperscript{e}</td>
<td>4.6167\textsuperscript{f}</td>
<td>2.8715\textsuperscript{f}</td>
</tr>
</tbody>
</table>

\textsuperscript{a} Pose RMSD calculated using PDB 4CI1 for comparison.
\textsuperscript{b} Pose RMSD calculated using PDB 4TZC for comparison.
\textsuperscript{c} Pose RMSD calculated using PDB 4TZ4 for comparison.
\textsuperscript{d} Pose RMSD calculated using PDB 4CI2 for comparison.
\textsuperscript{e} Pose RMSD calculated using PDB 4CI3 for comparison.
\textsuperscript{f} Pose RMSD calculated using PDB 4TZU for comparison.
**Figure 39.** Recombinant TBD protein expression and purification and zinc binding domain analysis using the zincon assay. (A) SDS-PAGE gel showing purification steps: A: total lysate, L: ladder, B: cleavage product, C: after 2nd GST purification, D: material loaded from size exclusion chromatography (SEC) with gel showing impurities, E-G: final samples after SEC. (B) Expression of recombinant TBD with Cys to Ser mutations. T is the total lysate and S is the soluble protein extract. (C) Structure of human TBD (PDB 4TZ4) in green, showing coordination of the zinc ion (red sphere) to the four cysteine residues. (D) Schematic of the zincon assay: zincon dye absorbs at 480 nm and upon chelation with zinc absorbs at 620-630 nm. (E) Wavelength scans showing a decrease at 480 nm and concomitant increase at 630 nm upon incremental addition of zinc. (F) Linear regression curve obtained from (E).
Figure 40. Binding affinity of IMiDs to TBD variants by ITC. ITC binding curves of pomalidomide (A-D), thalidomide (E-H), phthalimide (I-L) with wild-type and mutant CRBN TBD protein.
Figure 41. Binding pocket with N-methyl-lenalidomide
Figure 42. Drug Treatment of human and mouse T cells and multiple myeloma cell lines. MM1.S and 5TGM1 cells were treated with increasing concentrations of dBET1 for 24 hours. Protein expression levels of BRD4, c-Myc, CRBN, β-actin were determined by western blot.
Table 13. Statistical analysis of human T cells treated with IMiDs, JQ1, and dBET (for results in Figure 29).

<table>
<thead>
<tr>
<th>Relative Viability to DMSO (%)</th>
<th>Comparisons</th>
<th>P value</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>DMSO vs. JQ1 0.1</td>
<td>0.0036</td>
<td>**</td>
<td></td>
</tr>
<tr>
<td>DMSO vs. JQ1 1</td>
<td>0.0001</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>DMSO vs. JQ1 10</td>
<td>0.0001</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>DMSO vs. dBET 0.1</td>
<td>0.0001</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>DMSO vs. dBET 1</td>
<td>0.0001</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>DMSO vs. dBET 10</td>
<td>0.0001</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>DMSO vs. LEN</td>
<td>0.0794</td>
<td>ns</td>
<td></td>
</tr>
<tr>
<td>DMSO vs. POM</td>
<td>0.2335</td>
<td>ns</td>
<td></td>
</tr>
<tr>
<td>JQ1 0.1 vs. dBET 0.1</td>
<td>0.0004</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>JQ1 1 vs. dBET 1</td>
<td>0.0001</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>JQ1 10 vs. dBET 10</td>
<td>0.0001</td>
<td>***</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Relative c-Myc+ cells to DMSO (%)</th>
<th>Comparisons</th>
<th>P value</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>DMSO vs. Unstimulated</td>
<td>0.0001</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>DMSO vs. JQ1</td>
<td>0.0001</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>DMSO vs. dBET 0.1</td>
<td>0.0001</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>DMSO vs. dBET 1</td>
<td>0.0001</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>DMSO vs. dBET 10</td>
<td>0.0001</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>DMSO vs. dBET 50</td>
<td>0.0001</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>DMSO vs. M-dBET 0.1</td>
<td>0.3672</td>
<td>ns</td>
<td></td>
</tr>
<tr>
<td>DMSO vs. M-dBET 1</td>
<td>0.0777</td>
<td>ns</td>
<td></td>
</tr>
<tr>
<td>DMSO vs. M-dBET 10</td>
<td>0.0016</td>
<td>ns</td>
<td></td>
</tr>
<tr>
<td>DMSO vs. M-dBET 50</td>
<td>0.0001</td>
<td>***</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IL-2 Concentration (pg/mL)</th>
<th>Comparisons</th>
<th>P value</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>DMSO vs. Unstimulated</td>
<td>0.0001</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>DMSO vs. Len 10</td>
<td>0.0001</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>DMSO vs. POM 10</td>
<td>0.0001</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>DMSO vs. JQ1 0.1</td>
<td>0.2527</td>
<td>ns</td>
<td></td>
</tr>
<tr>
<td>DMSO vs. JQ1 10</td>
<td>0.0001</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>DMSO vs. dBET 0.1</td>
<td>0.9997</td>
<td>ns</td>
<td></td>
</tr>
<tr>
<td>DMSO vs. dBET 10</td>
<td>0.0002</td>
<td>***</td>
<td></td>
</tr>
</tbody>
</table>

*P value: * < 0.05, ** < 0.01, *** < 0.001, **** < 0.0001

Note: One-way ANOVA test and unpaired two tailed t-tests were performed.
Table 14. Statistical analysis of CRBN\textsuperscript{+/+} and CRBN\textsuperscript{−/−} mouse T cells treated with IMiDs, JQ1, and dBET (for results in Figure 30).

<table>
<thead>
<tr>
<th>Comparisons</th>
<th>Crbn\textsuperscript{+/+} T cells</th>
<th>Crbn\textsuperscript{−/−} T cells</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P value</td>
<td>Summary</td>
</tr>
<tr>
<td>DMSO vs. Unstimulated</td>
<td>0.0001</td>
<td>****</td>
</tr>
<tr>
<td>DMSO vs. Len 10</td>
<td>0.0033</td>
<td>**</td>
</tr>
<tr>
<td>DMSO vs. Pem 10</td>
<td>0.0606</td>
<td>ns</td>
</tr>
<tr>
<td>DMSO vs. JQ1 10</td>
<td>0.0001</td>
<td>****</td>
</tr>
<tr>
<td>DMSO vs. dBET 10.1</td>
<td>0.0001</td>
<td>****</td>
</tr>
<tr>
<td>DMSO vs. dBET 11</td>
<td>0.0001</td>
<td>****</td>
</tr>
<tr>
<td>DMSO vs. dBET 10</td>
<td>0.0001</td>
<td>****</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Comparisons</th>
<th>D96 MFI</th>
<th>C96 MFI</th>
<th>Relative Viability to DMSO (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P value</td>
<td>Summary</td>
<td>P value</td>
</tr>
<tr>
<td>DMSO vs. Unstimulated</td>
<td>0.0001</td>
<td>****</td>
<td>0.0001</td>
</tr>
<tr>
<td>DMSO vs. Len 10</td>
<td>0.0042</td>
<td>**</td>
<td>0.8548</td>
</tr>
<tr>
<td>DMSO vs. Pem 10</td>
<td>0.0166</td>
<td>*</td>
<td>0.8545</td>
</tr>
<tr>
<td>DMSO vs. JQ1 10</td>
<td>0.0001</td>
<td>****</td>
<td>0.0007</td>
</tr>
<tr>
<td>DMSO vs. dBET 10.1</td>
<td>0.0001</td>
<td>****</td>
<td>0.1391</td>
</tr>
<tr>
<td>DMSO vs. dBET 11</td>
<td>0.0001</td>
<td>****</td>
<td>0.9999</td>
</tr>
<tr>
<td>DMSO vs. dBET 10</td>
<td>0.0001</td>
<td>****</td>
<td>0.0003</td>
</tr>
</tbody>
</table>

P value: * < 0.05, ** < 0.01, *** < 0.001, **** < 0.0001

Note: One-way ANOVA test was performed.

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Division Index</th>
<th>CD96 MFI</th>
<th>Relative Viability to DMSO (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P value</td>
<td>Summary</td>
<td>P value</td>
</tr>
<tr>
<td>Unstimulated</td>
<td>0.155588997</td>
<td>0.58094322</td>
<td>0.83169971</td>
</tr>
<tr>
<td>DMSO</td>
<td>0.504643139</td>
<td>0.82919422</td>
<td>&gt;0.999999999</td>
</tr>
<tr>
<td>Len 10</td>
<td>0.196305179</td>
<td>0.306034149</td>
<td>0.01520159</td>
</tr>
<tr>
<td>Pem 10</td>
<td>0.585646279</td>
<td>0.340153344</td>
<td>0.02414454</td>
</tr>
<tr>
<td>JQ1 10</td>
<td>0.363603272</td>
<td>0.17599732</td>
<td>0.04173978</td>
</tr>
<tr>
<td>dBET 10.1</td>
<td>0.00690863</td>
<td>**</td>
<td>0.0013032</td>
</tr>
<tr>
<td>dBET 1</td>
<td>0.00725291</td>
<td>**</td>
<td>0.01198656</td>
</tr>
<tr>
<td>dBET 10</td>
<td>0.356298196</td>
<td>0.567427617</td>
<td>0.01302985</td>
</tr>
</tbody>
</table>
3.8 Chapter Appendix

3.8.1 Supplementary Computational Results

An in-depth theoretical investigation of the molecular binding mechanics of the immune-modulatory drugs (IMiDs) thalidomide, pomalidomide, and lenalidomide (Figure 43A-C) in complex with the protein cereblon (CRBN) was conducted to explore and understand possible differences in drug interactions between murine and human CRBN. Analysis of the published crystal structures for hCRBN, mCRBN, and gCRBN show negligible variation of RMSD between observed binding poses (Figure 43D and Figure 44F and G). To further test the structures for possible induced fit effects, protein flexibility, or crystal artifacts, MD simulations were performed to expand structural understanding of CRBN.

3.8.2 Determination of Equilibration for MD systems

Final system equilibration was determined by the observation of asymptotic behavior of the potential energy, RMSD, and Rg profiles and visual inspection of trajectories guided by RMSF profiles.

3.8.2.1 Equilibration of hCRBN model system. RMSD profiles of hCRBN in complex with DDB1 appear to show equilibration after approximately 25 ns with an average RMSD (Figure 35A) of 3.21 Å for the equilibrated region. Compared to the crystallographic structure resolution of 3.01 Å, the system RMSD is slightly outside the error of the reported PDB X-ray diffraction resolution. However, the RMSD profile of hCRBN residues associated with the known binding site (i.e. residues N335 to A421 and excluding DDB1) displays significant rigidity despite...
minimal asymptotic behavior portrayed by the RMSD graphs of the whole hCRBN chain. Further investigation using RMSF profiles (Figure 36I) finds that much of the displacement can be attributed to the solvent exposed N terminus tail region and from the DDB1 chain. RMSD profiles without the tail regions do show asymptotic behavior and the binding site structure is believed to be valid for the length of simulation. In a similar trend, Rg profiles (Figure 35E) do not appear to show equilibrated behavior for the hCRBN/DDB1 complex but do display clear asymptotic behavior after approximately 17 ns for hCRBN alone. The potential energy graphs (Figure 37M) also show asymptotic behavior after approximately 30 ns. From this data, all structures, analyses, and statistics for hCRBN are performed using the last 70 ns of the simulation and the stability and accuracy of the protein model of hCRBN are considered comparative to the original crystal structure (PDB 4TZ4) for the purposes of this study.
Figure 43. 3D chemical structures of IMiDs and CBN ribbon overlays. 3D chemical structures of thalidomide (A), lenalidomide (B), and pomalidomide (C). The S enantiomer was used for each compound. Ribbon overlays (D) of human (purple; PDB code 4TZ4) and chicken (blue; PDB code 4C12) X-ray crystal structures and post-MD structure of mouse (yellow; PDB code 4TZ4) cereblon. Human immunomodulatory drug–binding site residues His-378, Trp-380, Trp-386, and Phe-402 and ligands thalidomide (yellow), lenalidomide (green) and pomalidomide (blue) are shown for reference. Ribbon representations of the alanine mutation study (E) shown with the mutated residues H378A and W380A (green) of the post-MD haCRBN system (grey) juxtaposed with the WT residues of the original hCRBN system (purple).
Figure 44. Overlays with centered view of the CRBN thalidomide binding site. Images depict a superposition of backbones, represented as ribbons (F), and a superposition of co-crystallized ligands (H) of hCRBN (purple, PDB 4TZ4), gCRBN (blue, PDB 4CI2), and mCRBN (yellow, PDB 4TZ4) with residues H378, W380, W386, W400, F402, and the ligands thalidomide (yellow), lenalidomide (green), and pomalidomide (blue) shown for reference. Post-MD equilibrated systems of the CRBN thalidomide binding site after IFD displayed as a ribbon overlay (G) of hCRBN (red), hmCRBN (green), and gCRBN (orange), using the original crystal structure of hCRBN (purple, PDB 4TZ4) and lenalidomide (green) for reference. Superpositions of ligand poses (I) of lenalidomide (green) for the post-MD equilibrated systems of the CRBN thalidomide binding site after IFD for hCRBN (red), hmCRBN (green), gCRBN (orange), and the original crystal pose for hCRBN (purple) are also shown with the post-MD equilibrated protein structure of hCRBN (red) shown for reference.
3.8.2.2 Equilibration of mCRBN model system. The mCRBN system does not appear to be able to reach equilibrium. The average RMSD (Figure 35B) of approximately 3.27 Å is well outside the reported crystallographic structure resolution of 1.88 Å. The Rg profile (Figure 35F) and potential energy graph (Figure 37N) of the mCRBN simulation do not display asymptotic behavior. The RMSF profile (Figure 36J) and visual inspection reveals unfolding events of residues near the truncated regions of the protein which are the primary contributions to the average RMSD. However, the proximal residues 6 Å away from the ligand appear stable after 38 ns and may be able to be used for limited qualitative analysis. However, given the evidence provided, this model is not adequate for further modeling and analyses should be
viewed ephically. All structures, analyses, and statistics for mCRBN are performed using the last 62 ns of the simulation.

3.8.2.3 Equilibration of gCRBN model system. RMSD profiles of gCRBN in complex with DDB1 appear to show equilibration after approximately 32 ns with an average RMSD (Figure 35C) of 3.82 Å for the equilibrated region. Compared to the crystallographic structure resolution of 2.98 Å, the system RMSD is outside the error of the reported PDB X-ray diffraction resolution. However, in a similar trend to hCRBN, the RMSD profile of gCRBN residues associated with the known binding site (i.e. residues N335 to A421 and excluding DDB1) displays significant rigidity despite minimal asymptotic behavior portrayed by the RMSD graphs of the whole gCRBN chain. Further investigation using RMSF profiles (Figure 36K) finds that, again comparable to hCRBN, much of the displacement can be attributed to the solvent exposed N terminus tail region and from the DDB1 chain. RMSD profiles without the tail regions do show asymptotic behavior and the binding site structure is believed to be valid for the length of simulation. Rg profiles (Figure 35G) do not appear to show equilibrated behavior for the gCRBN/DDB1 complex but do display clear asymptotic behavior after approximately 22 ns for gCRBN alone. Despite potential energy graphs (Figure 37O) showing some continued declination, the error of the minimized fitted slope does encompass zero gradient after 7 ns. From this data, all structures, analyses, and statistics for gCRBN are performed using the last 27 ns of the simulation and the stability and accuracy of the protein model for the residues associated with the known binding site of gCRBN are considered comparative to the original crystal structure (PDB 4Cl2) for the purposes of this study.
3.8.2.4 Equilibration of hmCRBN model system. Equilibration can be safely assumed after 40 ns from the RMSD profiles of hmCRBN in complex with DDB1 with an average RMSD (Figure 35D) of 2.82 Å for the equilibrated region. Compared to the crystallographic structure resolution of 3.01 Å and the average RMSD calculated for the progenitor hCRBN system of 3.21 Å, the hmCRBN average system RMSD is well within error. Moreover, the RMSD profile of hmCRBN residues associated with the known binding site (i.e. residues N335 to A421 and excluding DDB1) also displays significant rigidity in confluence with the original hCRBN system. Further investigation of RMSF profiles (Figure 36L) finds, once again similar to the hCRBN system, that much of the displacement can be attributed to the solvent exposed N terminus tail region and from the DDB1 chain. Rg profiles (Figure 35H) do not appear to advance any evidence against equilibration and purport fluctuations directly analogous to the hCRBN system. The potential energy graphs (Figure 37P) also show asymptotic behavior after approximately 30 ns. From this data, all structures, analyses, and statistics for hCRBN are performed using the last 60 ns of the simulation and the stability and accuracy of the protein model of hCRBN are considered comparative to the original crystal structure (PDB 4C12) for the purposes of this study.

3.8.2.5 Equilibration of haCRBN model system. The RMSD (Figure 46A) and Rg (Figure 46B) plots for haCRBN do not show any asymptotic trend over the length of the simulation. Multiple distinct structural transitions are evident with the most pronounced conformational shift occurring after 8 ns when the binding site is observed to collapse, apparently due to the smaller occupied volume of the alanine mutation of residue 380 compared to tryptophan. The RMSD profiles also show a large degree of structural fluctuation through the
duration of the simulation. RMSF (Figure 46C) profiles also confirm this increased motion of
binding site residues compared to WT hCRBN (See graphical inlay of Figure 46C). Potential
energy graphs (Figure 46D) do display a negative slope, however, the error of the minimized
fitted slope does prelimit possible convergence after 12 ns. From the supplied evidence, this
model does not appear to converge toward equilibrium within the timespan of the simulation.
Even so, the simulation does provide some insights into the structural dependence of the CRBN
active site on the W380 residue and are discussed in the results section of the main text (Section
3.2.3 Immunomodulatory compound binding is conserved in CRBN sequence variants). All
structures, analyses, and statistics for haCRBN are performed using the last 82 ns of the
simulation with four explicit representative structures produced from hierarchical clustering. It
should also be noted that SiteMap did not find any suitable cavities near the original binding site
for any of these structures.
Figure 46. Equilibration profiles for haCRBN model system. (A) RMSD profiles of hCRBN. Profiles were created using backbone atoms and were calculated with hCRBN-DDB1 complex (Bottom Graph), hCRBN alone (Middle Graph), and binding site residues N335 to A421 (Top Graph). All measurement units are in Angstroms. (B) Rg Profiles of hCRBN. Profiles were created from all atoms and were calculated with hCRBN-DDB1 complex (Bottom Graph) and hCRBN alone (Top Graph). All measurement units are in Angstroms. (C) RMSF Profile of hCRBN. Profiles were created using backbone atoms and were calculated with hCRBN alone (DDB1 RMSF available upon request). Binding site residues N335 to A421 displayed as dashed line. All measurement units are in Angstroms. (D) Potential energy profile of hCRBN simulation. Multiple linear fits were characterized to determine asymptotic behavior by minimizing slope.

3.8.3 Determination of H378 protonation states

Determination of protonation states for protein side chains is an inherently difficult process. Crystallographic resolutions are rarely sufficient to resolve individual hydrogen atoms
and current methods for determining NMR coupling constants are not applicable to large proteins (330). This results in the need for indirect evidence to statistically characterize the ionization states for certain side chains. Hydrogen bonding analysis is a common practice of inferring protonation states (329). The determination of a hydrogen bond, however, is not a well defined parameter that largely succumbs to best practices. For this study, we use the metrics of a donor-acceptor distance of less than 3.5 Å and three point angle of greater than 90 degrees to quantify a hydrogen bond (327). For this study, careful attention must be made the residue H378 due to its involvement in the suspected binding motif of thalidomide. The six crystallographic structures of CRBN used in this study all appeared to show the H378 Nδ coordinated with the carbonyl oxygen of the ligand. This posits the assumption of a hydrogen bond and that the H378 Nδ must therefore be protonated. Unfortunately, studies have shown that reported structures do not necessarily portray the most stable isomer or statistically weight tautomeric states (329,331). For further insight into this issue, a separate system using a representative structure from the equilibrated hCRBN system was run to query alternate protonation states of this particular residue. PROPKA analysis before and after simulations were performed and confirm the assumption that the H378 Nδ is protonated. As such, all MD systems were generated with the neutral H378 δ-nitrogen as being protonated. PDB 4TZC and 4TZU (mCRBN) do not show the same interaction, however, as previously discussed, these systems do seem to have artifacts associated with crystal packing in which crystal mates may be interfering with the thalidomide binding site.
3.8.4 Computational Model Validation

All models developed through simulations, aside from mCRBN and haCRBN, are statistically valid for reproducing physical results and support the physical relevance of the crystal structures (See Section 3.8.2 Determination of Equilibration for MD Systems). While the mCRBN crystal structure is valid, it is not sufficient for computational modeling (See Section 3.8.2.2 Equilibration of mCRBN Model System). This is probably due to truncation of the protein sequence and the omission of DDB1. Crystal packing effects appear to provide surrogate stability for the absent DDB1 cofactor, allowing for crystallization but limiting the effectiveness of computational modeling on mCRBN alone. This artificial instability of the native structure likely imposes a manufactured mutability of the thalidomide binding site as evidenced by poor reproduction and comparison of IFD results shown in Table 12.

The murine CRBN analog, hmCRBN, developed from the hCRBN system appears to be capable of reproducing the crystal ligand poses of PDB 4TZC and 4TZU (mCRBN) and the equilibrated hmCRBN system displays minimal conformational deviation from the murine crystals (Figure 44G and Figure 45K). The hmCRBN system can therefore be considered a suitable substitute for mCRBN for modeling purposes. Further, the three residue distinctions between hCRBN and mCRBN do not appear to have any observable relevance in the structure or function of CRBN. The E377 residue for hCRBN does not appear to interact with the ligands in any significant way. This is evidently due to backbone dihedral strain which forces the charged carboxyl moiety away from the binding site and minimizes possible interactions with the ligands (Figure 45K). The unexpected absence of this interaction helps to explain why the binding modes do not appear to differ between hCRBN and mCRBN. Binding modes also do not appear to differ between models and compounds as evidenced by negligible differences in RMSD
calculations between crystal binding poses and post-MD models (Figure 44F-I and Figure 45J). IFD also predicts no observable difference in binding affinity between models and compounds (Table 12). All poses are within 1.8 Å which is the expected threshold reported by Sherman et al. for the IFD protocol (123).

3.8.5 CP Consensus Docking Results

A CP consensus docking study was performed on hCRBN with thalidomide, lenalidomide, and pomalidomide (S enantiomers only) as known literature controls, SG5-003, MANT uracil, CC-122, and Glutarimide as positive project controls, and uracil, phthamimide, NCI373535, and SG5-005 as negative controls and decoys. Docking was performed on the prepared, MD equilibrated hCRBN model (Section 3.8.2.1 Equilibration of hCRBN model system). The results of the analysis are shown in Table 15.
Table 15 gives a clear breakdown of computational error and protein model characteristics, allowing for a simple rescaling of docking results using Pearson correlations as regression weights and confidence in relying only on XP docking for additional analog testing (Table 16). From Table 6, we can infer several properties of the hCRBN binding site and binding modes from CP consensus docking results. Due to low correlation to IFD and
MM/GBSA and high correlation to XP, ligand binding to CRBN TBD does not appear to have any induced fit effects. Since the binding site is not catalytic, binding must interfere with or induce alternate protein-protein interactions. Further, high correlation to XP and low correlation to SP indicate a significantly hydrophobic pocket with hydrogen bonding donor and acceptors in the hydrophobic enclosure. Strong correlation to QPLD also suggests a complex electronic environment in the binding site, understandable due to the aromatic cage that creates the TDB pocket. Poor correlation to MM/GBSA also represents little dependence on solvent. This is a reasonable assumption since there are no apparent structural waters in or near the binding site and the pharmacophore occludes any space for solvent to intrude. From these propositions, SiteMap analysis, and ligand structural overlays, a simple pharmacophore can be developed to fully explain the binding mechanism of hCRBN (Figure 45J). Additionally, docking results can now be scaled to fit the absolute binding energies from experiment, allowing predictive computational measurement for novel analogs with optimal efficiency. The higher range of docking scores fit better with experiment, therefore, subsequent docking scores are normalized with a bias toward less negative energies.

3.8.6 High Accuracy Course-Grained Docking Derived from CP Consensus Docking

The complete pharmacophore coupled with high correlations of XP docking with controls provides ample evidence that a course-grained virtual screening model using only XP docking can reproduce physical results to a sufficient degree of accuracy compared to experiment and high resolution modeling. Table 16 is an example of applying this course-grained model to analogs. The drug discovery project was able to drastically reduce ITC and fluorescence experiments, reducing costs and speeding up production. Positive and negative controls should
be included to scale expected binding affinities to the controls. **Table 16** employs a conservative binary ranking of “possible” or “unlikely” estimates of a hit for analogs. The distribution of binding affinities for the controls could reasonably allow for up to four categories of affinity estimates.

### 3.8.7 Conclusions from Modeling

Due to truncation of the protein sequence and the omission of DDB1, mCRBN is not a sufficient model for virtual study, leading to poor reproduction and comparison of IFD results. The murine CRBN analog, hmCRBN, however, can be used as a surrogate for mCRBN. Alanine mutations of hCRBN, haCRBN, show the significant residues forming the aromatic cage of the CRBN-TBD. When the residues are mutated to the smaller alanine side chain, the binding site collapses after a few nanoseconds of MD. All other models show excellent agreement with experimental crystal structures, pose RMSD, and estimated free energies of binding. This indicates the exceptional ability of the designed computational models to reproduce and predict physical quantities.

Further, ITC experiments for $K_D$ determination used a racemic mixture of each IMiD control. Docking runs on each enantiomer exhibited marginally better binding of the S enantiomer compared to the R enantiomer (**Table 16**). This finding lines up well with recent experiments able to overcome the optical inversion of IMiD enantiomers. Mori et al. demonstrated a significant difference in enantiomer $K_D$, $3.5 \pm 0.4 \mu M$ for (S)-thalidomide and $20.0 \pm 2.9 \mu M$ for (R)-thalidomide (336). Re-scaled computational results are within experimental error and confirm the analytical and predictive power of the CP consensus docking methodology.
Table 16. Course-grained virtual screening model utilizing only XP docking for assessing analog binding affinity.

<table>
<thead>
<tr>
<th>Compound</th>
<th>XP Docking Score</th>
<th>Likelihood of Binding</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>NL-M-01</td>
<td>-12.151</td>
<td>Possible</td>
<td></td>
</tr>
<tr>
<td>Lenalidomide</td>
<td>-11.105</td>
<td>Possible</td>
<td>Control: Known Binder (Green)</td>
</tr>
<tr>
<td>Lenalidomide-R</td>
<td>-11.057</td>
<td>Possible</td>
<td>Control: Enantiomer (Yellow) - Lower Score Than S (Good)</td>
</tr>
<tr>
<td>Pomalidomide</td>
<td>-11.006</td>
<td>Possible</td>
<td>Control: Known Binder (Green)</td>
</tr>
<tr>
<td>Thalidomide</td>
<td>-10.916</td>
<td>Possible</td>
<td>Control: Known Binder (Green)</td>
</tr>
<tr>
<td>SG5-001</td>
<td>-10.562</td>
<td>Possible</td>
<td></td>
</tr>
<tr>
<td>CC-122</td>
<td>-10.495</td>
<td>Possible</td>
<td>Control: Known Binder (Green)</td>
</tr>
<tr>
<td>SG-M-06</td>
<td>-10.273</td>
<td>Possible</td>
<td></td>
</tr>
<tr>
<td>SG-M-05</td>
<td>-10.218</td>
<td>Possible</td>
<td></td>
</tr>
<tr>
<td>Thalidomide-R</td>
<td>-10.157</td>
<td>Possible</td>
<td>Control: Enantiomer (Yellow) - Lower Score Than S (Good)</td>
</tr>
<tr>
<td>SG-M-03</td>
<td>-10.126</td>
<td>Possible</td>
<td></td>
</tr>
<tr>
<td>SG-M-10</td>
<td>-10.114</td>
<td>Possible</td>
<td></td>
</tr>
<tr>
<td>SG-M-09</td>
<td>-10.041</td>
<td>Possible</td>
<td></td>
</tr>
<tr>
<td>SG-M-11</td>
<td>-10.014</td>
<td>Possible</td>
<td></td>
</tr>
<tr>
<td>SG5-068</td>
<td>-9.984</td>
<td>Possible</td>
<td></td>
</tr>
<tr>
<td>SG-M-02</td>
<td>-9.956</td>
<td>Possible</td>
<td></td>
</tr>
<tr>
<td>SG-M-12</td>
<td>-9.954</td>
<td>Possible</td>
<td></td>
</tr>
<tr>
<td>SG-M-04</td>
<td>-9.899</td>
<td>Possible</td>
<td></td>
</tr>
<tr>
<td>NL-M-02</td>
<td>-9.843</td>
<td>Possible</td>
<td></td>
</tr>
<tr>
<td>SG5-003</td>
<td>-9.766</td>
<td>Possible</td>
<td>Control: Known Binder (Green)</td>
</tr>
<tr>
<td>MANT uradil</td>
<td>-9.687</td>
<td>Possible</td>
<td>Control: Known Binder (Green)</td>
</tr>
<tr>
<td>SG-M-08</td>
<td>-9.629</td>
<td>Possible</td>
<td></td>
</tr>
<tr>
<td>NL-M-03</td>
<td>-9.621</td>
<td>Possible</td>
<td></td>
</tr>
<tr>
<td>MA7-026</td>
<td>-9.304</td>
<td>Possible</td>
<td></td>
</tr>
<tr>
<td>MA6-174-9</td>
<td>-9.233</td>
<td>Possible</td>
<td></td>
</tr>
<tr>
<td>SG-M-07</td>
<td>-9.194</td>
<td>Possible</td>
<td></td>
</tr>
<tr>
<td>Pomalidomide-R</td>
<td>-9.189</td>
<td>Possible</td>
<td>Control: Enantiomer (Yellow) - Lower Score Than S (Good)</td>
</tr>
<tr>
<td>MA6-180-3</td>
<td>-8.977</td>
<td>Unlikely</td>
<td></td>
</tr>
<tr>
<td>Guanamiridic</td>
<td>-8.575</td>
<td>Unlikely</td>
<td>Control: Known Average Binder (Orange)</td>
</tr>
<tr>
<td>NL-M-04</td>
<td>-8.132</td>
<td>Unlikely</td>
<td></td>
</tr>
<tr>
<td>SG5-061</td>
<td>-7.827</td>
<td>Unlikely</td>
<td></td>
</tr>
<tr>
<td>MA7-032</td>
<td>-7.48</td>
<td>Unlikely</td>
<td></td>
</tr>
<tr>
<td>SG5-092</td>
<td>-7.325</td>
<td>Unlikely</td>
<td></td>
</tr>
<tr>
<td>SG5-078</td>
<td>-7.258</td>
<td>Unlikely</td>
<td></td>
</tr>
<tr>
<td>MA6-174-1</td>
<td>-7.219</td>
<td>Unlikely</td>
<td></td>
</tr>
<tr>
<td>Uracil</td>
<td>-7.169</td>
<td>Unlikely</td>
<td>Control: Known Poor Binder (Red)</td>
</tr>
<tr>
<td>NGC373535</td>
<td>-6.878</td>
<td>Unlikely</td>
<td>Control: Known Poor Binder (Red)</td>
</tr>
<tr>
<td>SG5-003</td>
<td>-6.435</td>
<td>Unlikely</td>
<td>Control: Known Poor Binder (Red)</td>
</tr>
<tr>
<td>Phthamiride</td>
<td>-6.154</td>
<td>Unlikely</td>
<td>Control: Known Poor Binder (Red)</td>
</tr>
<tr>
<td>R-SG-M-01</td>
<td>-5.108</td>
<td>Unlikely</td>
<td></td>
</tr>
<tr>
<td>SG5-093</td>
<td>-5.333</td>
<td>Unlikely</td>
<td></td>
</tr>
</tbody>
</table>

Experimentally determined strong binders in green, average binders in orange, and poor binders in red. IMiD enantiomers in yellow.
4 Discovery of a Non-Nucleotide Potential STING Antagonist from Consensus Docking and Dimer Identification with Novel Site-Restiction Virtual Screening

4.1 Introduction

STING (Stimulator of Interferon Genes), also known as MITA, MPYS, and ERIS, encoded by gene TMEM173, is a facilitator of the innate immune system through sensing cyclic dinucleotides created by cGAS after interacting with cytosolic DNA and RNA. STING then promotes production of cytokines and type 1 interferons, α and β, by propagating the signal by means of interferon regulatory factor 3 (IRF3) phosphorylation (337). This is the first report on the discovery of STING as an essential innate immune regulator (338,339). This report demonstrates that CDNs bind to and activate STING. This invasion of aberrant cytosolic DNA can be caused by various infectious agents including bacterial and viral infection (340). Other sources can be from self-DNA breaching nuclear confines due to improper cell division or DNA damage caused by oncogenic processes. Self-DNA present in cytosol may also come from some autoimmune disorders such as systemic lupus erythematosus (SLE) or Aicardi–Goutières syndrome (AGS) (341). This leads to a surprising number of diseases and disorders that intersect the STING pathway and gives rise to potentially novel therapeutic options for these difficult to treat diseases.

Recently, there has been interest in developing agonists to increase activity of the STING pathway as a modality for cancer treatment (342). The idea is to stimulate the immune system to target tumor cells. Most STING agonists developed to date have been cyclic di-nucleotides
(CDNs). An exception is a recently reported amidobenzimidazole STING agonist (343). A CDN agonist drug candidate from Aduro Biotech, Inc. has entered clinical trials (344). However, no STING antagonists have entered clinical trials.

STING is a promising new target for systemic lupus erythematosus (SLE) therapy due to its response to cytosolic DNA. The cytosolic enzyme cGAS (cyclic GMP-AMP synthase) is a surveillance enzyme that detects cytosolic DNA and produces a molecule referred to as 2’3’-cGAMP (cyclic [G(2’,5’)pA(3’,5’)p]). 2’3’-cGAMP activates the STING pathway through direct binding to STING (Figure 47). This pathway has been proposed as a cause of increased expression of IFIT3 and other type 1 interferons in SLE (345–347). STING, an endoplasmic reticulum bound cytoplasmic protein, is key in the cellular defense when there is cytosolic DNA or RNA. Blocking STING activation may prevent anti-DNA antibody production in SLE.

There is a dire need for new drugs to treat SLE since many drug candidates are failing in clinical trials. In 2018, only one of the eight drug candidates in mid to late trials attained significant efficacy and that candidate required higher doses to reach efficacy so safety issues must be readdressed (347). Anti-DNA autoantibodies are a hallmark of SLE, occur in most SLE patients, and can appear years prior to SLE symptoms (348). Rather than stimulating STING to provoke an immune anti-tumor response, the objective of a STING antagonist would be to suppress anti-DNA antibody production that might attack endogenous DNA in some autoimmune disease models, allowing treatment options at an earlier interdiction point in the disease.

Since DNA normally resides in the mitochondria, its presence in the cytoplasm causes a danger-associated molecular pattern (DAMP). STING works through the mechanism of nucleic acids from bacteria or viruses inducing immune responses in cells (349). Once free-floating exogenous DNA or RNA is identified, CGAS is the sensor that detects DNA as a DAMP, making
cyclic dinucleotides. Cyclic dinucleotides such as cyclic di-guanine monophosphate (c-di-GMP) and cyclic [G(2′,5′)pA(3′,5′)p] (2,3-cGAMP) are immune stimulators which regulate type I interferon (IFN-α and β) induction (350). These special nucleotides are secondary messenger molecules which bind to STING, a protein localized on the endoplasmic reticulum membrane. Binding of cyclic dinucleotides induces a conformational shift in STING allowing the recruitment of tank-binding kinase 1 (TBK1) and transcription factor IRF3 to STING (Figure 47) (351). IRF3 then localizes to the nucleus forming a transcription factor complex with NFkB, AP-1, P65, P50, and CBP (352).

NF-kB and IRF3 are components of the IFN-β promoter enhanceosome that forms to stimulate the IFN-β gene transcription. The enhanceosome (composed of ATF2 and c-JUN, IRFs, NF-kB, and HMG-I(Y)) recruits histone acetyl transferases (HATs) to locate the transcription start site acetylating a subset of lysine residues in the histones of the nucleosome. This ultimately recruits the TFIID transcription complex to the IFN-β gene promoter (353). IFN-β secretion engages the type I IFN receptor to activate JAK-STAT (signal transducer and activator of transcription) signal transduction and tyrosine phosphorylation of STAT1 and STAT2 and the transcription of their target genes (354). The IFN-β promoter is very weakly associated with TNF-a in that TNF-a can activate the IFN-β promoter which leads to activation of NF-kB and AP-1 but not IRFs (355).

Two homologous proteins, IRF3 and IRF7 are key regulators of type I IFN gene expression induced by viruses. IRF3 resides latently in the cytosol. It is constitutively expressed and undergoes phosphorylation, dimerization, and nuclear translocation upon viral infection (356).
Figure 47. STING pathway. To understand the direct binding mechanism of STING to other protein binding partners, Sitemap surface modeling was carried out to determine the most likely alignment between the ordered lid region of STING with TBK1. Sitemap surface analyses also showed significant similarities between the surface of IRF3 and the ordered lid region of STING indicating that TBK1 likely binds to both in the manner shown above. This information was needed to determine what residues are important for binding and what protein conformations were essential for propagating the signal downstream. From structural data using PDB 4LOH, 4IM0 and 5JEJ, a more comprehensive diagram of STING interactions with TBK1 and IRF3. TBK1 and IRF3 facilitate IRF3 phosphorylation, leading to IRF3 dimerization and translocation to the nucleus. Upon translocation, the IRF3 dimer activates IFN-β gene transcription. IFN-β is responsible for antiviral, antibacterial and anticancer properties.

In 2011, Vance et al. discovered that cyclic diguanylate monophosphate (c-di-GMP) binds to mammalian STING receptors, activating an innate immune response (339). Then, Sun et al. found that an enzyme called cyclic GMP-AMP synthase (cGAS) binds DNA enabling two nucleotides to covalently bond together creating a cyclic dinucleotide (CDN) (357). The prospect of using manufactured CDNs or other compounds which modulate STING have caused many
biopharma companies around the world to extensively study STING for the treatment of cancer, infectious diseases as well as use in vaccines as adjuvants.

Early studies on STING focused mainly on mSTING due to its high sequence homology (~89%). The compound DMXAA arose as a promising lead compound from these initial experiments, even though it ultimately failed in human trials (358). Despite the lack of efficacy in human trials, we compared the crystal structures between mSTING and hSTING and found that, while the two forms have significant homology and structural similarities, the SiteMap analyses of the CDN binding site show that they have completely different binding site topologies (Figure 48). Therefore, mSTING was not deemed necessary for inclusion in this study. This observation may also explain differences in IFN-β fold induction among compounds between the two species.

Leticia Corrales of Aduro biotech published the groundbreaking paper that lead the field towards its current, feverish state of intense research (344) Concerns of toxicity and an uncontrolled “cytokine storm” lead to reevaluation of STING agonists upon revelation of the dark side of STING upregulation (359,360). While there are many types of STING agonists, cyclic di-nucleotides (CDNs) are the most prominent type of molecule being proposed to upregulate the activity of STING. Examples of such molecules can be found (361). These agonists mimic the structure and therefore activity of 2,3-cGAMP in the binding pocket of STING. While very potent, these ubiquitous prokaryotic intracellular signaling molecules lack efficacy due to their need for intratumoral injection (IT) which prohibits treatment of non-cutaneously accessible tumors (344) as well as requiring multiple injections and concomitant administration of immune checkpoint antibodies (362). Clinical trial NCT02675439 which investigates IT of a CDN as monotherapy may prove to be insufficient to treat patients with
advanced/metastatic solid tumors and lymphomas. Clinical trial NCT03010176 uses a CDN in conjunction with a checkpoint inhibitor to treat patients with advanced/metastatic solid tumors and lymphomas. Through biomarker evaluation of patients, a non-inflamed could become and T-cell inflamed tumor microenvironment which would presumably make patients more sensitive to checkpoint inhibitor therapy. The current objective response rate to anti-PD-1 therapy is about 40–45% in patients with PD-L1-positive tumors (363–365) (Merck, NCT03010176).

The STING cytosolic pathway was demonstrated to be involved in cellular response to diverse viral and bacterial infections with cytomegalovirus being one of the most widespread viral agents among human populations. STING was implicated in sensing and controlling its replication in vitro by means of activation of cGAS-STING-TBK1-IRF3 pathway (366,367). Additionally, Reinert et al. demonstrated that cGAS-STING signaling pathway is extensively involved in response to herpes simplex virus infection among microglia via activation of IFN I through TLR3 pathway, and effects of innate immune priming. Furthermore, cytosolic DNA sensing pathway- deficient mice experienced elevated receptiveness to herpes simplex encephalitis (368). STING was also found to be associated with cytokine production in response to Legionella pneumophila bacterial infection. STING or cGAS deficiency in murine model and HAQ TMEM173/STING allele expression indicated a strong predisposition to the bacterial infection; mice with impaired cGAS-STING pathway experienced higher bacterial loads, compared to the control groups. Ruiz-Moreno et al. explicitly correlated HAQ allele expression with advanced bacterial and viral infection rates (369).

On the other hand, STING agonists shown by (370) activate STING, leading to immunostimulatory downstream effects. This study showed that CDNs injected without delivery agents diffused away rapidly which did not appear to cause any lasting cytotoxic effects. Another
STING agonist, ML RR-S2 CDA shows promise in clinical trial NCT02675439 due to its efficacy in promoting the specific rejection of several types of tumors in murine models. (AduroBiotech, NCT02675439)

Figure 48. SiteMap analysis between CDN binding sites of hSTING (B) and mSTING (A). Hydrogen bond acceptor regions shown in blue, hydrogen bond donor regions shown in red, and hydrophobic regions shown in yellow. Residues R238, Q266, and T267 are displayed for reference. (C) Ribbon overlay comparing structures of mSTING (light blue) and hSTING (orange).
4.2 Experimental Design and Methods

4.2.1 Computational Methods

4.2.1.1 Protein Preparation. Protein model systems hSTING variants are prepared using the Schrodinger software suite (25). Protein structure coordinates are downloaded from the Protein Data Bank (PDB) (227,228). The hSTING models are generated from the PDB entries (Table 17): 4LOH (REF allele co-crystallized with 2’3’-cGAMP), 4LOI (REF allele co-crystallized with 2’2’-cGAMP) (371), 4EMT (REF allele co-crystallized with c-di-GMP), 4EMU (REF allele apo structure) (349), 4KSY (WT allele co-crystallized with 2’3’-cGAMP) (372), 4F5Y (WT allele co-crystallized with c-di-GMP) (373), and 4F5D (HAQ allele co-crystallized with c-di-GMP) (373). An MD simulation was also performed on the hSTING$_{WT}$ structure with 2’3’-cGAMP bound mutated to the hSTING$_{HAQ}$ isoform using PDB 4KSY to simulate 2’3’-cGAMP structural effects on the HAQ variant. The apo structures for hSTING, PDB 4EMU, is used to cross reference conformational states from MD simulations.

Table 17. List of PDB entries used for MD simulation.

<table>
<thead>
<tr>
<th>PDB ID</th>
<th>hSTING Allele</th>
<th>Binding Site Residue Positions</th>
<th>Co-crystallized Ligand</th>
</tr>
</thead>
<tbody>
<tr>
<td>4LOH</td>
<td>REF</td>
<td>G 230 H 232</td>
<td>2’3’-cGAMP</td>
</tr>
<tr>
<td>4LOI</td>
<td>REF</td>
<td>G 230 H 232</td>
<td>2’2’-cGAMP</td>
</tr>
<tr>
<td>4EMT</td>
<td>REF</td>
<td>G 230 H 232</td>
<td>c-di-GMP</td>
</tr>
<tr>
<td>4EMU</td>
<td>REF</td>
<td>G 230 H 232</td>
<td>-</td>
</tr>
<tr>
<td>4KSY</td>
<td>WT</td>
<td>G 230 R 232</td>
<td>2’3’-cGAMP</td>
</tr>
<tr>
<td>4F5Y</td>
<td>WT</td>
<td>G 230 R 232</td>
<td>c-di-GMP</td>
</tr>
<tr>
<td>4F5D</td>
<td>HAQ</td>
<td>A 230 R 232</td>
<td>c-di-GMP</td>
</tr>
</tbody>
</table>

207
PDB systems are prepared with the Protein Preparation Wizard (PrepWizard) in Maestro (231,232). Cofactors used in crystallization (such as sulfate or phosphate ions), ligands, and additional protein dimers are deleted. Bond orders are then assigned, including disulfide bridges, and original hydrogens are deleted and later replaced to reduce bad contacts and other crystal artifacts before protonation and hydrogen bond optimization. All waters are retained for assisting in the determination of side chain protonation states and initial hydrogen bond optimization. Missing side chains are added and optimized using Prime (16). Hydrogen atoms are then added to the protein, remaining cofactors, and to any added structural waters. The program PROPKA (50) is used for the prediction of protein ionization states at 7.4 pH and ProtAssign is used for hydrogen bond optimization. After automatic hydrogen assignment, visual inspection is used to flip residues and change protonation states at protein-protein interfaces if and when appropriate.

4.2.1.2 Molecular Dynamics. MD simulations (Table 17) are performed with the GPU accelerated Desmond MD program (233–236) on two Nvidia GeForce GTX 1080 Ti video cards. A cubic simulation box is created extending at least 10Å from the protein with imposed periodic boundary conditions using TIP3P waters (237) as solvent. The OPLS-3 all-atom force field (34) is then applied to all atoms. Simulations are run at a temperature of 310 K and a constant pressure of 1 atm. All systems are energy minimized followed by multiple restrained minimizations to randomize systems before equilibration and final simulation. Production MD is performed on all systems for 250 ns.

Final system equilibration is determined by the observation of asymptotic behavior of the potential energy, Root Mean Square Deviation (RMSD), and Radius of Gyration (Rg) profiles
and visual inspection of trajectories guided by Root Mean Square Fluctuation (RMSF) profiles (Figure 57).

4.2.1.3 Consensus Docking. After equilibration is determined, a hierarchical average linkage clustering method based on RMSD was utilized to determine an average representative structure for the equilibrated hSTING systems. The program PROPKA is then implemented again on the equilibrated structure to test consistency of side chain protonation states at 7.4 pH. The representative structure is then used for consensus docking incorporating five diverse and complementary docking methods, SP and XP rigid receptor docking, Induced Fit Docking, Quantum Polarized Ligand Docking, and MM/GBSA methods. Each described in detail below. By applying these varied energy scoring methods, the weaknesses of each method can be identified for a particular model and error statistically minimized, yielding a more accurate summary of ligand binding dispositions and affinities.

As a check for the placement of the grids used in the docking studies and for further analysis of the binding cavity for the CDN binding site, Schrödinger's SiteMap program (245–247) is employed. SiteMap searches the protein structure for likely binding sites and highlights regions within the binding site suitable for occupancy by hydrophobic groups, hydrogen-bond donors, acceptors, or metal-binding functionality of the ligand. All ligands were prepared using the program LigPrep (248) and the OPLS-3 all-atom force field (34) was applied to all ligand atoms.

4.2.1.4 Rigid Receptor Docking (RRD). Rigid docking simulations were performed by the docking program Glide (114,121,125,335). Glide uses a GlideScore fitness function based
on Chemscore (249,250) for estimating binding affinity, but includes a steric-clash term, adds buried polar terms to penalize electrostatic mismatches, and modifies other secondary terms. Docking simulations used both the standard precision (SP) and extra precision (XP) methods. XP mode is a refinement algorithm enforced only on good ligand poses. Sampling is based on an anchor and refined growth strategy and the scoring function includes a more complete treatment of some of the SP energetic terms, such as the solvation and hydrophobic terms. Docking grids were defined by a rectangular ligand atom inclusion outer box of 22Å and ligand centroid constraint inner box of 10Å in the x, y, and z directions originating from the binding cavity centroid defined by SiteMap.

4.2.1.5 Induced Fit Docking (IFD). The IFD methodology (122–124,334) incorporates both the docking program Glide to account for ligand flexibility and the Refinement module in the Prime program to account for receptor flexibility. The Schrödinger IFD protocol attempts to model induced-fit effects from alterations in binding site conformation due to ligand binding in order to increase accuracy of binding affinity estimates and prediction of possible binding modes.

Separate cubic docking grids for the CDN binding site are centered on the original co-crystallized ligand centroid and from the binding cavity centroids defined by SiteMap. A constrained minimization of the receptor is performed preceding an initial softened potential Glide docking of the ligand set is then implemented with the standard precision (SP) mode. The resulting top 20 poses of the ligands are used to sample protein plasticity by conformational searches and minimizations of binding pocket residues within 6 Å of any ligand pose for all
complexes obtained. The new receptor conformations are then redocked using complexes within 30 kcal/mol from the best scoring structure.

The estimated binding affinity of each complex is reported in the GlideScore and used to compare differences between each ligand while the Emodel score is used to inter-compare poses of the ligands. Emodel places more significance on weighting force field components (electrostatic and van der Waals energies), making it better for comparing conformers as opposed to comparing chemically-distinct species.

**4.2.1.6 Quantum Polarized Ligand Docking (QPLD).** To account for ligand polarization upon binding, Quantum Mechanics / Molecular Mechanics (QM/MM) docking is performed by the Schrödinger QM-Polarized Ligand Docking Protocol (QPLD) (126,127). The protocol first employs RRD using Glide in SP mode. In this step, the top five poses of each ligand in the initial RRD are used. Potential ligand polarization induced by the protein are then calculated with Qsite (111,251,252) at the B3LYP/6-31G* level. The ligand force fields are then reconstructed with QM/MM modified charges, redocked, and five poses of each ligand are saved for evaluation.

**4.2.1.7 Molecular Mechanics and Generalized Born Surface Area (MM/GBSA).** The MM/GBSA method combines molecular mechanics energy terms and implicit solvation models to calculate the binding-free energy based on docking complexes. The protocol, implemented by the Prime MM-GBSA module, calculates optimized free energies for the free protein and free ligand and references them with the original bound complex energy (253). Polar contributions are calculated using the Generalized Born (GB) model (254), an implicit solvent model is based
on a variable dielectric surface Generalized Born (VD-SGB) approach, where the variable
dielectric value for each residue was fit to a large number of side-chain and loop predictions
while the non-polar energy is estimated using the solvent accessible surface area (SASA) (255).
The simulation was performed based on receptor–ligand complex structures obtained from
induced fit docking. The obtained ligand poses were minimized using the local optimization
feature in Prime, whereas the energies of complex were calculated with the OPLS-3 force field
and Generalized-Born/Surface Area continuum solvent model (256). During the simulation
process, the ligand strain energy is also considered.

4.2.2 Experimental Methods

4.2.2.1 Surface Plasmon Resonance (SPR). Surface Plasmon Resonance (SPR) was
employed for binding measurements using His-tagged hSTING\textsuperscript{WT} CDN domain. A GE
Healthcare Biacore T200 was equipped with an Ni-NTA chip. 16,951 RU of 6X-His tagged
human STING was crosslinked via NHS chemistry following injections of 350 mM EDTA and
500 mM NiSO\textsubscript{4}. STING natural substrates and the lead compound were titrated and flowed at 60
uL/min in 1X PBS for 60 sec association time followed by a 135 sec dissociation. The
sensorograms were analyzed using Biacore T200 Software 3.0 (GE Healthcare) and steady state
was measured at 4 sec before injection stop, exported into Graphpad, and fit vs concentration
using a one site specific binding model to calculate the apparent equilibrium dissociation
constant (K\textsubscript{D}). Where appropriate, kinetics were measured using a 1:1 Langmuir binding model
with R\textsubscript{max} set to local to obtain the association rate (K\textsubscript{on}), dissociation (K\textsubscript{off}), and the K\textsubscript{D}.
4.2.2.2 Microscale Thermophoresis (MST). Nanotemper Monolith NT.115 labeled thermophoresis machine was used with standard treated capillary tubes using samples comprised of labeled protein and titrations of small molecule in 1X PBS. MST experiments were conducted in triplicate mixing 200 nM protein with 100 nM dye and allowing to sit at room temperature for 30 minutes followed by centrifugation on Ni-NTA 488 labeled His-labeled hSTING\textsuperscript{WT}. Detection of the protein was performed using the blue detection channel with LED excitation power set to 90% and MST set to high allowing 3 s prior to MST on to check for initial fluorescence differences, 25 s for thermophoresis, and 3 s for regeneration after MST off. Analysis was performed using M.O. Affinity Analysis Software with difference between initial fluorescence measured in the first 5 s as compared with thermophoresis at 15 s at 15 different analyte concentrations ranging from 15 nM to 1 mM and exported into Graphpad Prism v.8 using a Log inhibitor v. response 4 parameter fit.

4.2.2.3 Luciferase Assay. THP1-ISG-Lucia cells were obtained from Invivogen and maintained in RPMI 1640 containing 2 mM L-glutamine, 25 mM HEPES, 10% heat-inactivated fetal bovine serum, 100 μg/ml Normocin™, Pen-Strep (100 μg/ml). To maintain selection pressure, 10 μg/ml of blasticidin and 100 μg/ml of Zeocin™ was added to the growth medium every other passage.

Reporter cells were plated at 100,000 (THP1-ISG-Lucia) cells per well in a white 300uL sterile 96 well plate and treated with 3, 5, 10, 15, 20, 30, 40, 50μM 2’3’-cGAMP, DMSO and NSC 335504 (Figure 53). For compounds, 10mM stock solution in 100% DMSO was diluted 1:4 with ultrapure Milli Q water. For the positive control, 1mM stock solution in 100% Milli Q H₂O was diluted 3:1 with water. 50μL of QUANTI-Luc™ luminescence assay reagent (Invivogen,
San Diego, CA, USA) was added after 18-24h incubation period. Expression of Lucia luciferase was quantified by measuring luminescence from duplicate treatments. Data illustrated are average luminescence changes shown relative to DMSO-treated cells.

Using a Glowmax luminometer (Promega, Madison, WI, USA). Prepare QUANTI-Luc™ following the manufacturer’s instructions. Pipet 10 µl of THP1-Dual™ KI-hSTING-R232 cell culture supernatant per well in a 96-well white (opaque) plate. Add 50 µl of QUANTI-Luc™ per well. Proceed immediately with the measurement using a 4sec incubation time and integrating over 1 sec.

THP1-Dual KI-hSTING-R232 Cells from Invivogen contain a knockin of the intronless coding sequence of the R232 hSTING variant. This variant, which contains an arginine at position 232 (R232), is the most prevalent variant (~45-58%) in the human population (344). This isoform is preferentially activated by 2’5’linkage-containing cGAMP isomers (339).

4.2.2.4 Chemical Synthesis of NSC335504. NSC335504 (Clonixeril) was synthesized by employing substituted nicotinic acid esters (or free acids) and substituted anilines. Figure 49 shows the full synthesis scheme utilized to make the compound.
4.3 Results & Discussion

4.3.1 Computational Agonist and Antagonist Model Construction

Initial computational studies employed extensive MD simulations to better understand how STING interacts with ligands and other potential binding partners. Some commonalities seem to arise from comparing trajectories between isoforms and control ligands leading to postulation on unique and intricate binding mechanics. The main binding partners for STING in the IFN-β pathway are IRF3 and TBK1 (351). A simple metric was devised to measure possible activation of STING by evaluating the distance between the alpha carbons of residue H185 at the end of the α2 helix (Figure 51). Crystal structures of known agonists present alpha carbon distances in the range of 34 to 38 Å. Apo crystal structures have alpha carbon distances in the range of 47 to 56 Å (Table 18). The proposed binding mechanism is that initial binding is with the bottom of the CDN binding region near residues Q266 and T267. Then, as proposed by
several authors (372–374), the initially disordered “lid” region comprising residues 154-244 (Figure 51, green ribbon) interacts with the bound ligand and induces a β sheet formation, bringing the α2 helices closer together. The ligand then associates itself more with the ordered lid region, further stabilizing the conformation.

This realization allowed the ability to form two separate ensemble docking models to screen for agonists, compounds that have greater affinity to the holo structure and thus stabilizing the ordered lid allowing for signaling to continue, and for antagonists or partial agonists in which compounds would have better affinity with the disordered apo structure, limiting the lid microfolds from occurring and thus attenuating further downstream signaling. Experimental evidence provides solid support for these theoretical models. ITC experiments from Zhang et al. (372) give evidence that the binding mechanism for 2’3’-cGAMP is largely entropy driven (-TΔS = 12.199 kcal/mol; ΔH = 0.71 kcal/mol) denoting decreasing entropy from reordering of the lid region upon binding, while c-di-GMP is mostly enthalpic with a large entropy penalty (-TΔS = -12.117 kcal/mol; ΔH = -20.19 kcal/mol). This large entropy gain when binding c-di-GMP clearly shows that STING is more disordered when binding to this compound but still energetically favorable due to better protein-ligand contact from induced fit effects with a more flexible protein structure. SPR binding kinetics also lend support for these separate mechanisms. The k_{off} rates for 2’3’-cGAMP and c-di-GMP differ by nearly an order of magnitude, k_{off} = 9.72x10^{-4} and 8.76x10^{-3} respectively. Since 2’3’-cGAMP does not have better enthalpic interactions than c-di-GMP, quite the opposite from the difference in ΔH, the longer ligand association time must, therefore, be due to significant induced folding.

Phenotypic prediction based on protein dynamics may also be possible by determining the H185 alpha carbon distance after MD simulation with the associated ligands (Figure 58).
Measured distances for crystal structures and 250 ns MD simulations (Table 18) showed excellent systematic agreement for 2’3’-cGAMP (35 Å for crystal (372), 37.8 Å for MD structure), an artificial analog 2’2’-cGAMP (38.4 Å for crystal (371), 41.9 Å for MD structure), c-di-GMP (41.4 Å for crystal (349), 46.1 Å for MD structure), and apo structures (47.1 Å for crystal (349), 55.5 Å for MD structure, Figure 59). RMSD profiles of MD trajectories fit well with crystallographic B-factors overall (Figure 57) and MD models appear to better resolve binding energetics and comparisons to literature values.

4.3.2 STING Isoform Selection

Additional complexities arise when accounting for the different STING alleles. Profiles of RMSD per residue from MD simulations show significant contrasts between the REF, WT, and HAQ isoforms. The lid regions for REF and WT are substantially more disordered than HAQ (Figure 50). This may be due to increased entropy from the glycine residue at the 230 position for REF and WT, opposed to alanine for HAQ (Figure 51). This could account for some discrepancies between experimental binding energies and model calculations as current computational techniques cannot easily measure this.
Figure 50. RMSF profiles between REF (green), WT (red), and HAQ (blue) isoforms of STING bound to 2′3′-cGAMP. Residues for lid region shown as dashed lines.
**Figure 51.** hSTING<sup>REF</sup> CTD structures after MD with distance between H185 residue shown for holo structure bound to 2’3’-cGAMP, red carbons, (left) and apo structure (right). “Lid” region demarcated with green ribbon and α1 and α2 helix in light blue. Residues H185, G230, and H232 represented as tubes for clarity.

To account for potential differences between STING isoforms, computational models were created using only those isoforms which had amino acid variations in the known STING binding site, residues 230 and 232. This lead to three main models, hSTING<sup>WT</sup> (G230, R232), hSTING<sup>REF</sup> (G230, R232H), and hSTING<sup>HAQ</sup> (G230A, R232). Technically, there is no difference between AQ and HAQ structure in terms of binding site residues, but for consistency, we will continue to reference the G230A R232 variant as the HAQ structure. No literature data was found for ITC K<sub>D</sub> values on control compounds (2’3’-cGAMP, 2’2’-cGAMP, and c-di-GMP) for the HAQ variant. Since these values are imperative for proper model validation, hSTING<sup>HAQ</sup> was not used for comparing model controls with literature values. We performed ligand binding analyses using five different binding energy estimation algorithms (SP, XP, IFD, QPLD, and MM/GBSA) to evaluate theoretical binding energies of control compounds to the hSTING<sup>WT</sup> and hSTING<sup>REF</sup> isoforms to determine the optimal docking algorithm and default isoform. Statistical models were compiled based on deviation from known values and internal variance to adjust for
docking and simulation error. ITC, SPR, and consensus docking energetics for the hSTING\textsuperscript{WT} native ligands 2’3’-cGAMP, 3.8 nM (372), 1.4 nM, 2.4 nM respectively, and c-di-GMP, 1.2 μM (372), 4.8 μM, and 6.4 μM respectively (Table 18), agree with unequivocal precision, lending substantial evidence to the validity of the hSTING\textsuperscript{WT} binding models. Due to the WT allele dominance in human populations and the accuracy of docking replication of controls, the MD equilibrated hSTING\textsuperscript{WT} antagonist (PDB 4F5Y) and hSTING\textsuperscript{WT} agonist (PDB 4KSY) conformation were then taken as the two generalized conformer docking models for STING.

Table 18. Model comparisons to literature and experiment.

<table>
<thead>
<tr>
<th>Structure</th>
<th>Compound</th>
<th>H185 Distance (Angstroms)</th>
<th>H185 Post-MD Distance (Angstroms)</th>
<th>Model $K_d$ (nM)</th>
<th>SPR $K_d$ (nM)</th>
<th>ITC $K_d$ (nM)</th>
<th>Cell EC\textsubscript{50} (nM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>hSTING\textsuperscript{WT}</td>
<td>c-2’3’-GAMP</td>
<td>35.0\textsuperscript{a}</td>
<td>37.8</td>
<td>2.4</td>
<td>1.4</td>
<td>3.8\textsuperscript{a}</td>
<td>42 (IFNβ mRNA)\textsuperscript{a}</td>
</tr>
<tr>
<td></td>
<td>c-2’2’-GAMP</td>
<td>-</td>
<td>43.1</td>
<td>256</td>
<td>-</td>
<td>287\textsuperscript{a}</td>
<td>16 (IFNβ mRNA)\textsuperscript{a}</td>
</tr>
<tr>
<td></td>
<td>c-di-GMP</td>
<td>53.0\textsuperscript{a}</td>
<td>56.5</td>
<td>6377</td>
<td>4776</td>
<td>1210\textsuperscript{a}</td>
<td>538 (IFNβ mRNA)\textsuperscript{a}</td>
</tr>
<tr>
<td>hSTING\textsuperscript{REF}</td>
<td>c-2’3’-GAMP</td>
<td>34.7\textsuperscript{b}</td>
<td>39.3</td>
<td>784</td>
<td>-</td>
<td>5300\textsuperscript{b}</td>
<td>1200 ELISA\textsuperscript{b}</td>
</tr>
<tr>
<td></td>
<td>c-2’2’-GAMP</td>
<td>38.4\textsuperscript{b}</td>
<td>41.9</td>
<td>236</td>
<td>-</td>
<td>2500\textsuperscript{b}</td>
<td>3400 ELISA\textsuperscript{b}</td>
</tr>
<tr>
<td></td>
<td>c-di-GMP</td>
<td>52.6\textsuperscript{d}</td>
<td>54.1</td>
<td>1300</td>
<td>-</td>
<td>4600\textsuperscript{e}</td>
<td>ND (IFNβ Luc)\textsuperscript{c}</td>
</tr>
</tbody>
</table>

\textsuperscript{a} Zhang et al. (372), \textsuperscript{b} Gao et al. (371), \textsuperscript{c} Shu et al. (349), \textsuperscript{d} Ouyang et al. (375), \textsuperscript{e} Shang et al. (373). ND (Not detected). Cell assay type given in parentheses.
4.3.3 Consensus Docking Results

The consensus docking approach employed in this work attempts to apply docking protocols that specialize in determining pose prediction and scoring from more encompassing physical properties (e.g. polarizabilty, receptor flexibility, solvent interaction, etc.) and weighting the output of each simulation to yield more insight into individual protein-ligand complexes. Consensus docking performed in this manner can partition the physical characteristics of a binding pocket to help elucidate the binding mechanisms of that site. With appropriate controls, an optimal course-grained algorithm can also be identified for virtual screening of larger compound libraries. Docking was performed on the prepared, MD equilibrated hSTING\textsuperscript{WT} antagonist (PDB 4F5Y) and hSTING\textsuperscript{WT} agonist (PDB 4KSY) models.
Table 19. Consensus docking using disparate docking methodologies to partition physical characteristics of the STING-CDN binding site.

<table>
<thead>
<tr>
<th>Compound</th>
<th>Combined Ensemble</th>
<th>Agonist Model (Post-MD)</th>
<th>Antagonist Model (Post-MD)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>XP</td>
<td>SP</td>
<td>IFD</td>
</tr>
<tr>
<td>2′,3′-cGAMP</td>
<td>-10.802</td>
<td>-8.324</td>
<td>-12.531</td>
</tr>
<tr>
<td>2′,2′-cGAMP</td>
<td>-7.121</td>
<td>-9.488</td>
<td>-11.594</td>
</tr>
<tr>
<td>c-di-GMP</td>
<td>-6.233</td>
<td>-5.801</td>
<td>-10.179</td>
</tr>
<tr>
<td>NSC335504 (LD)</td>
<td>-7.860</td>
<td>-8.782</td>
<td>-10.614</td>
</tr>
<tr>
<td></td>
<td>Average</td>
<td>+/− Error (SD)</td>
<td>Theoretical K_D (nM)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pearson</td>
<td>0.974</td>
<td>0.327</td>
<td>0.922</td>
</tr>
<tr>
<td>R²</td>
<td>0.948</td>
<td>0.107</td>
<td>0.851</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MM/GBSA Error (SD)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Experimental K_D values were converted to ΔG° for R² and Pearson correlations to docking runs. NSC335504 docking scores are for the zero-order bond linked dimer (LD).

Table 19 gives a clear breakdown of computational error and protein model characteristics. Docking correlations between ensembles provide further evidence of the need for separate docking models to assess conformational stabilization of the active or inactive forms of STING. R² and Pearson correlations of control compounds with experimentally determined K_D values (Table 18) show poor correspondence when docked to only one conformational model. However, when the ensembles are combined based on which model best approximates experimental values, correlations show extremely reliable agreement to controls. From this, we can infer several properties of the hSTINGWT binding site and binding modes. Low correlation to
SP docking and underestimated binding affinities of XP docking coupled with higher correlations to flexible docking methods (IFD and MM/GBSA) strongly suggest that ligand binding is dependent on significant induced fit effects and domain conformational shifts. This is obvious from the distinct conformational differences of the apo and holo crystal structures, though reveals a potential binding mechanism of antagonistic molecules. Antagonistic molecules not only search for the initial open conformation of the apo structure, but likely induce and stabilize a disordered lid conformation while maintaining a strong affinity for the site. Additionally, high correlation and accurate estimation of control ligand affinity through QPLD implies the site has a complex electrostatic environment capable of repolarizing susceptible ligands. This is a reasonable assumption owing to the presence of highly charged residues in the site, depending on the isoform tested. It may be due to the lack of charge characteristic and increased flexibility of the lid region that the REF isoform does not bind or produce as much of a cellular response to control ligands. Coupling IFD and MM/GBSA methods appears to balance out affinity estimations. IFD generally overestimates affinities, while MM/GBSA shows consistently lower affinity values. As the site is largely exposed to solvent, dependence on solvent interactions is presumably significant and, therefore, understandable that the Generalized Born and Surface Area continuum solvation algorithm would correlate well with controls.

From these propositions, SiteMap analysis, simulations, and reproduction of literature values, a general pharmacophore can be developed to explain the binding mechanism of hSTING. Additionally, after comparing multiple docking algorithms between each model structure, docking results can now be scaled to fit the absolute binding energies from experiment, allowing predictive computational measurement for novel analogs with optimal efficiency. XP docking was found to be a sufficient coarse screen method using the hSTING$^{WT}$ agonist and
antagonist models. The NCI diversity set V and ASINEX diversity sets were screened and a coarse-grained pharmacophore model was employed to rapidly scan the ZINC database using the ZINC Pharmer algorithm (376). Hits from ZINC Pharmer were then rescored through XP docking. Top hits were reviewed and best candidates were chosen for experimental testing to confirm hits.

4.3.4 Site-Restriction Virtual Screening for Identification of Possible Dimer Complexes

The STING-CDN binding site is a relatively voluminous pocket with a complex and symmetrical binding surface between the two protein monomers. Hence, there is significant potential for a small molecule to form a dimeric complex inside the binding site. Current docking programs are bounded to docking and evaluating a single molecule at a time and are, thus, unable to effectively interpret ligand dimerization. To overcome this limitation, we developed a simple docking method that can be used in conjunction with standard virtual screening protocols to assist in identifying potential dimer complexes.

Following typical compound screening of the whole site (Figure 52B), an additional screening is then performed with the ligand restricted to a monomeric half of the binding site (Figure 52A). This procedure, as opposed to only docking to the protein monomer alone, ensures the site electrostatics are consistent between the whole site and the restricted portion. The next step is to implement a RMSD comparison of the ligand poses for the whole and half site docking runs. If the RMSD between the whole and half site poses is less than the commonly accepted 2 Å cutoff, then the docked ligand evidently prefers a specific region in the binding site and should allow for another stoichiometric equivalent of the molecule to bind into the surplus volume.
After review of the initial ligand poses, updated docking grids are generated with the original docked compound and the unoccupied region subsequently screened with a duplicate ligand (Figure 52C). Providing the postliminary screen is successful at producing a reasonable pose, the dimer composite structure can then be linked through a zero-ordered bond connecting the two most proximal atoms. For Glide, this type of bond only has an enforced distance constraint, angle and dihedral terms are zero, and does not interfere with the molecular force field. Docking the linked dimer (LD) back into the respective conformer will allow the docking algorithm to properly calculate estimated free energies of binding for the LD-protein complex.

Figure 52. Site-restriction docking method for identifying potential dimeric ligand complexes. Initial docking to monomer unit of binding site is performed with half of the site excluded (A). Secondary re-docking of ligand is the performed with no restrictions (B). If ligand maintains pose in both docking simulations, a second copy of the ligand is re-docked to a new grid with the original ligand held in place (C). The re-docked copy is then linked to the initial pose with a zero-order bond, connecting the most proximal atoms, and re-docked once more (D, molecular surface shown for clarity). Lid region demarcated with green ribbons. Green boxes indicate ligand centroid positional constraint and purple boxes represent all ligand atom positional constraint. Antagonist model with NSC335504 shown as reference.
From this screening protocol, the compound NSC335504 (Figure 53) was identified as a potential hit (Figure 52D). Consensus docking was implemented on the dimer structure with a predicted affinity range of 70 nM to 5.7 μM. Further experimental assays were performed for verification.

![NSC 335504](image)

**Figure 53.** Molecular representation of NSC335504 (left) and associated mass spectrum profile (right). NSC335504 has a molecular weight of 336.772 g/mol and a molecular formula of C_{16}H_{17}ClN_{2}O_{4}.

### 4.3.5 Direct Binding Assays

Computational modeling predicted an average binding affinity $K_D$ of 627 nM for NSC335504. Microscale Thermophoresis (MST) was performed as a rough initial estimate of binding affinity, yielding a $K_D$ of 260 ± 66 nM and SPR was then used to refine and confirm results with a $K_D$ of 430 ± 140 nM (Figure 54, left and right respectively). SPR was also
implemented on the native ligands c-di-GMP and 2’3’-cGAMP (Figure 55). Direct binding assays are well within agreement with literature values and computational models (Table 18 and 19). Therefore, computational modeling can be heavily relied on for lead optimization for either agonists, partial agonists, or antagonists with the NSC335504 scaffold.

Figure 54. MST (Left) and SPR (Right) binding assays for NSC335504. SPR steady state utilized hill model yielding a Hill coefficient determined as 1.738, indicating ligand dimerization near 2:1 binding.

Figure 55. Steady state SPR binding affinity plots for c-di-GMP and 2’3’-cGAMP controls.
4.3.6 THP-1 Luciferase Reporter Assay

In order to determine the cellular activity of NSC335504, a luciferase reporter assay of IRF3 expression was employed to qualitatively determine activation of the STING pathway by NSC335504. The cell-based assay exhibited insignificant upregulation of IRF3 in comparison to 2,3-cGAMP (Figure 56).

Modeling and direct binding analysis clearly indicate NSC335504 strongly binds to the STING receptor (Table 18 and Figure 54) but does not appear to activate the STING pathway with any significant potency. This evidence combined with modeling data of NSC335504 favoring an inactive conformation of STING strongly suggests the potential antagonistic behavior of NSC335504.

NSC335504 is the National Cancer Institute (NCI) accession number for clonixeril, the glyceryl ester of clonixin. Formerly used as an NSAID for pain and inflammation, particularly in transdermal drug delivery formulations (377), we believe the clonixeril structure has been under-utilized, with possibilities for more specific targeting.
4.4 Conclusions

MD equilibrated crystal structures for human HAQ, REF, and WT alleles were clustered to find optimal conformations for diverse chemical library screening. Novel consensus docking protocols utilizing rigid receptor, induced fit, and quantum polarized ligand docking were applied for quantifying and refining proposed binding mechanisms of STING isoforms. Models for STING agonists and antagonists were developed and rigorously tested against literature and experimental biochemical and cellular studies.

From directed virtual screening, a novel low-molecular-weight organic molecule, NSC335504, not based on a cyclic dinucleotide was found as a potential STING deactivator and is currently under investigation. Our hypothesis is that NSC335504 acts as an antagonist, potentially competing with 2,3-cGAMP. SPR and MST characterize strong binding, while a
luciferase assay displays negligible STING activation compared to the native ligand suggesting an antagonistic mechanism.

Small molecule antagonists described in (378,379) demonstrate that our approach to inhibiting STING is viable. The study provides support for the efficacy of STING antagonists in the treatment of autoinflammatory disease. The progress of this compound will depend on multiple validation studies and the modifications of the compound to create a more bioavailable compound. Future projects will focus on developing a competition assay to confirm NSC335504 as an antagonist and synthesizing analogs from the NSC335504 scaffold guided through rational drug design.
4.5 Supplementary Material

4.5.1 Supplemental Figures
Figure 57. RMSD and Radius of Gyration plots for MD systems. RMSD and $R_g$ profiles of hSTING$^{REF}$ are as follows: 4LOH (A), 4LOI (B), 4EMT (C), and 4EMU (D). RMSD and $R_g$ profiles of hSTING$^{WT}$ are as follows: 4KSY (E) and 4F5Y (F). RMSD and $R_g$ profiles of hSTING$^{HAQ}$ are as follows: 4F5D (G) and 4KSY G230A mutation (H). Profiles were created using backbone atoms and all measurement units are in Angstroms.

Figure 58. H185 Ca distances for MD systems. Structures for hSTING$^{WT}$ are as follows (Solid lines): Dark Red – 2’3’-cGAMP, Dark Blue – 2’2’-cGAMP, Dark Green – c-di-GMP. Structures for hSTING$^{REF}$ are as follows (dashed lines): Light Red – 2’3’-cGAMP, Light Blue – 2’2’-cGAMP, Light Green – c-di-GMP.
Figure 59. H185 Ca distance for STING$^{\text{REF}}$ apo structure MD system.
References


analysis of macromolecular motions in a database framework: developing mode concentration as

71. Delarue M, Dumas P. On the use of low-frequency normal modes to enforce collective
movements in refining macromolecular structural models. Proc Natl Acad Sci. 2004 May
4;101(18):6957–62.

72. Kolossváry I, Guida WC. Low Mode Search. An Efficient, Automated Computational
Method for Conformational Analysis: Application to Cyclic and Acyclic Alkanes and Cyclic

73. Mahajan S, Sanejouand Y-H. On the relationship between low-frequency normal modes
and the large-scale conformational changes of proteins. Arch Biochem Biophys. 2015 Feb
1;567:59–65.


75. Rathore N, de Pablo JJ. Monte Carlo simulation of proteins through a random walk in

Molecular Biology™). Available from: https://doi.org/10.1007/978-1-59745-177-2_2

77. Release S. MacroModel. Schrödinger LLC N Y NY. 2016;


79. Hatch HW, Mahynski NA, Shen VK. FEASST: Free Energy and Advanced Sampling


99. Head-Gordon M, Rico RJ, Oumi M, Lee TJ. A doubles correction to electronic excited
states from configuration interaction in the space of single substitutions. Chem Phys Lett. 1994

100. Møller Chr, Plesset MS. Note on an Approximation Treatment for Many-Electron


102. Mardirossian N, Head-Gordon M. How Accurate Are the Minnesota Density Functionals
for Noncovalent Interactions, Isomerization Energies, Thermochemistry, and Barrier Heights
Involving Molecules Composed of Main-Group Elements? [Internet]. 2016 [cited 2019 Mar 28].
Available from: https://pubs.acs.org/doi/abs/10.1021/acs.jctc.6b00637

103. Paier J, Marsman M, Kresse G. Why does the B3LYP hybrid functional fail for metals? J


105. Costa HFM da, Trsic M, Mohallem JR. Universal gaussian and Slater-type basis sets for
atoms He to Ar based on an integral version of the Hartree-Fock equations. Mol Phys. 1987 Sep
1;62(1):91–5.


107. Gronert S. The need for additional diffuse functions in calculations on small anions: the

245


249. Murray CW, Auton TR, Eldridge MD. Empirical scoring functions. II. The testing of an empirical scoring function for the prediction of ligand-receptor binding affinities and the use of


257. Page CS, Bates PA. Can MM-PBSA calculations predict the specificities of protein
258. Gohlke H, Case DA. Converging free energy estimates: MM-PB(GB)SA studies on the
Cε Regulates γ-Aminobutyrate Type A Receptor Sensitivity to Ethanol and Benzodiazepines
structure–activity relationship of 3,4′-bispyridinylethylenes: Discovery of a potent 3-
isoquinolinylpyridine inhibitor of protein kinase B (PKB/Akt) for the treatment of cancer. Bioorg
Suppressor ARID1A Controls Global Transcription via Pausing of RNA Polymerase II. Cell Rep.


289. Gandhi AK, Kang J, Havens CG, Conklin T, Ning Y, Wu L, et al. Immunomodulatory agents lenalidomide and pomalidomide co-stimulate T cells by inducing degradation of T cell...


268


272


352. Jones C. Regulation of Innate Immune Responses by Bovine Herpesvirus 1 and Infected Cell Protein 0 (bICP0). Viruses. 2009 Sep;1(2):255–75.


Appendices

Contained herein are further works, included as reprints, during graduate studies outside the immediate preview of this opus and the associated documents for previously published material.
Appendix 1: Mixing Properties of Sphingomyelin Ceramide Bilayers: A Simulation Study
Mixing Properties of Sphingomyelin Ceramide Bilayers: A Simulation Study

Rainer Metcalf and Sagar A. Pandit

Department of Physics, University of South Florida, Tampa, Florida 33620, United States

ABSTRACT: Ceramide is the simplest molecule in the class of glycosphingolipids composed of a sphingosine backbone and acyl moiety. It plays significant roles in cell signaling; apoptosis; binding of hormones, toxins, and viruses; and many other biologically important functions. Sphingomyelin, ceramide with a phosphatidylcholine headgroup, is another biologically vital lipid present in the myelin sheath of nerve cell axons. Regions with high concentrations of ceramide can be formed in biological membranes composed of sphingomyelin by enzymatic catalysis with sphingomyelinas. To better understand the biophysical and thermodynamic properties of these molecules and their mixtures, we have performed NPT molecular dynamics simulations of hydrated 16:0 sphingomyelin bilayers with increasing concentrations of 16:0 ceramide at 323, 332, 340, and 358 K. From analyses of electron densities, hydrogen bonding, NMR order parameters, partial molecular volume, and partial molecular area, we have identified possible structural changes corresponding to liquid ordered and liquid disordered phases. These structural changes are the results of changes in intra- and intermolecular hydrogen bonds between SM and Cer molecules. Our results correspond to DSC experiments for sphingomyelin bilayer concentrations up to 50% Cer. Above 50% concentration, we observe conformational changes in the SM headgroup similar to that of the umbrella model for lipid cholesterol mixtures.

INTRODUCTION

Ceramide (N-acyl α-hydroxylated sphingosine) has become a considerable topic of study after it was discovered to be an important signaling molecule. The biological significance of ceramide (Cer) includes its roles in the binding of hormones, toxins, and viruses; as secondary messengers in signaling pathways; and in apoptosis. Cer is derived from the aliphatic amino alcohol sphingosine ([2S,3R,5E]-2-amino-4-octadecene-1,3-diol) and is the hub of sphingolipid metabolism. It is present in relatively diffuse amounts in cell membranes, except for the stratum corneum, where the largely nonpolar molecule also serves as a structural component in membranes. Cer decreases permeability in these membranes, effectively “waterproofing” the epidermis. Cer-rich ordered domains of increased order and rigidity are also implicated in many other diverse biological functions.

Sphingomyelin (SM) is another type of glycosphingolipid closely related to ceramide. SM consists of a Cer unit with an esterified phosphorylcholine moiety. The amine-linked acyl chains are usually saturated or monounsaturated and contain 14–24 carbons. SM is a major component of the myelin sheath that surrounds nerve axons. These membranes act as insulators, facilitating action potential across the nerve membrane by inhibiting ion dissipation into surrounding media. Demyelinating diseases such as multiple sclerosis, Guillain–Barré syndrome, and leukodystrophy are thought to originate from the deterioration of these membranes.

Ceramide can be created from the degradation of SM. This reaction is mediated by the enzyme sphingomyelinase, which causes a hydrolysis reaction with the phosphoester bond of SM, yielding phosphocholine and Cer. Some potent inducers of sphingomyelinase activity are proinflammatory cytokines, such as tumor necrosis factor-α (TNF-α), interleukin-1β (IL-1β), interferon-γ (IFN-γ), and bacterial lipopolysaccharides. Studies done by Singh et al. have also found that reactive oxygen species, such as hydrogen peroxide, can cause degradation of SM membranes in astrocytes, oligodendrocytes, microglia, and glial cells. This leads to the creation of significant quantities of Cer, 18–25% sphingomyelin conversion to Cer, that accumulate in the membrane, leading to death of the cells. Clearly, Cer and SM are important biological molecules.

Cellular metabolism is significantly affected by the structural properties of bilayers composed of these lipids. NMR experiments by Huang et al. have shown that the addition of Cer to dipalmitoylphosphatidylcholine (DPPC) membranes at 45 °C induced lateral phase separation of the bilayers into regions of gel and liquid crystalline phases. The author reasoned this effect was due to packing defects in the bilayer. NMR studies performed by Thewalt et al. also showed that complex mixing...
behavior arose when Cer was introduced to sphingomyelin bilayers. It was suggested that,\textsuperscript{14,23,24} despite being structurally different from cholesterol (Chol), the rigidity of long-chain Cer is similar to the rigidity of the sterol rings of Chol. This combination gives rise to detergent-resistant membrane regions of increased lipid chain order. Xu et al.\textsuperscript{25–27} observed that Cer stabilizes domain formation in SM and other phosphocholine lipid mixtures with Chol. It was also noted that Cer displaces Chol from those domains. Further, the authors suggest that tight lipid chain packing is essential for sterol displacement by Cer. This packing behavior also prevents unfavorable contacts between lipid headgroups and water. Recent studies have also shown that, depending on the Chol/Cer ratio, Chol can displace Cer, as well, and influence the generation and distribution of SM/Cer domains.\textsuperscript{28} Deuterium NMR studies by Hsueh et al.\textsuperscript{29} found that domains of gel and liquid-crystalline phases coexist in mixtures of palmitoyloleoyl phosphocholine (POPC) and SM over a wide range of temperature and compositions above 1S mol % Cer. These experiments give rise to the idea that the effect of Cer on lipid chain packing in low concentrations exhibits properties similar to that of Chol.\textsuperscript{30}

Simulations consisting of these lipids and their mixtures are useful in elucidating underlying interactions between Cer and SM. Insights gleaned from these interactions can then be connected to their biological activity. Molecular dynamics (MD) simulations performed by Chiu et al.\textsuperscript{31} showed the dynamics of SM systems are distinctly different from those of a DPPC bilayer. The authors also remarked on the relative lack of experimental and computational studies done on these biologically important mixtures of SM. Zhang et al.\textsuperscript{32} suggested that phase separation was closely linked to the balance of energy and entropy change when mixing SM, POPC, and Chol. Pandit et al.\textsuperscript{33} proposed through MD simulations that Cer may, under some circumstances, act as a surrogate for Chol and may also be an evolutionary precursor to it.\textsuperscript{34} As described above, because of the ease of accumulation of Cer via enzymatic catalysis of SM, comparisons of structural effects of Cer on SM bilayers are important in understanding the mechanisms of some diseases and the biological effects of these molecules and their mixtures.

**METHODS**

The MD simulations were performed on hydrated 16:0 SM bilayers with varying concentrations of 16:0 Cer ranging from 0 to 100% at 10% increments. In addition, simulations with 5% concentrations of Cer were also performed for better sampling of concentration. Figure 1A and B show the structure of the Cer and SM molecules respectively, along with the associated atom names. The force-field parameters used for the SM and Cer molecules were taken from previous work.\textsuperscript{35,36} Our systems were simulated at four different temperatures—323, 332, 340, and 358 K—using the Nose–Hoover temperature coupling scheme.\textsuperscript{37} Each system consisted of 200 lipids with appropriate proportion of SM and Cer molecules. The systems were hydrated with 201000 SPC/E water molecules. This ensured a lipid water ratio of 1:100. This translates to ~71.9 wt % water for pure SM and ~83.7 wt % water for pure Cer. The amount of water used in these simulations was adequate because a Cer bilayer is considered fully hydrated at 74.3 wt % water or a lipid water ratio of 1.86.\textsuperscript{38}

These simulations were performed using the GROMACS package, version 3.3.\textsuperscript{39} The linear constraint solver LINCS algorithm was used to constrain all bonds in the systems,\textsuperscript{39} allowing an integration time step of 2 fs. Periodic boundary conditions were applied in all three dimensions, and long-range electrostatics were calculated using the SPME algorithm\textsuperscript{40} with a real-space cutoff of 9.5 Å. A cutoff of 16 Å was employed for van der Waals interactions. The systems were simulated in an NPT ensemble using the Parrinello–Rahman pressure-coupling scheme at a constant pressure of 1 atm.\textsuperscript{40}

Initial configurations for all the mixture systems were generated by random placement of 100 bilayer molecules per leaflet in appropriate proportion such that the lipids were aligned along the z axis and the hydrocarbon chains were pointing toward the z = 0 plane. Two blocks of 10,000 SPC/E waters were placed above and below the constructed bilayers. The systems were energy-minimized to remove bad contacts resulting from overlapping hard sphere surfaces and over stretched bonds.

All the systems were annealed to ensure proper thermalization of the hydrocarbon chains. The annealing steps involved raising the initial temperature of the systems to 600 K and lowering it by steps of 50 K per 250 ps until the systems reached the desired temperatures. After the annealing process, 30 ns continuous MD simulations were performed on each system. The box areas and volumes were monitored throughout the duration of the simulations. All analyses were performed on the last 10 ns of each

---

**Figure 1.** Structures of palmitoyl 16:0 Cer (A) and palmitoyl 16:0 SM (B). Molecules shown with atom names.
Table 1. Table of Comparisons for Pure SM Systems

<table>
<thead>
<tr>
<th>quantity</th>
<th>temperature, K</th>
<th>results</th>
<th>Niemela et al. ( ^{43} )</th>
<th>Chiu ( ^{et al.} ) ( ^{34} )</th>
<th>Mehner et al. ( ^{46} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>area per lipid (Å²)</td>
<td>323</td>
<td>53.3 ± 0.4</td>
<td>52 ± 1</td>
<td>53</td>
<td></td>
</tr>
<tr>
<td>volume per lipid (Å³)</td>
<td>323</td>
<td>1167.8 ± 12.1</td>
<td>1110 ± 10</td>
<td>1182</td>
<td></td>
</tr>
<tr>
<td>( \delta_{\text{H}} )</td>
<td>323</td>
<td>43.8 ± 0.3</td>
<td>43.4 ± 0.5</td>
<td>42.4</td>
<td>0.221</td>
</tr>
<tr>
<td>( \delta_{\text{D}} )</td>
<td>323</td>
<td>0.246</td>
<td>0.26</td>
<td>0.25</td>
<td></td>
</tr>
</tbody>
</table>

*Simulations performed with 18:0 SM.

simulation where fluctuations in the box areas and volumes were minimal.

Pandit \( ^{33} \) have reported the inaccuracies in measurements of volumetric properties of mixed bilayers for short time scale simulations (i.e., less than 50 ns). These inaccuracies stem from the changes due to slow lateral molecular diffusion in mixed lipid bilayers. However, for the simulations reported here, the area per molecule and lateral mean square displacement begin to equilibrate around 20 ns. From the previous author’s data, this yields a possible error in the area per molecule. Our analyses are based on a comparison between systems, and therefore, any systematic error is expected to change quantitative results by up to a constant. Thus, these simulations provide accurate qualitative description of lipid mixing. Much larger systems and longer run times may be necessary to obtain a more comprehensive quantitative measure of lipid mixing.

Results

Validation of the Simulations. The simulations were validated first by comparing the structural properties of the simulated systems with the experimental and previously reported simulation studies. The systems with 0% and 100% Cer were used for this purpose.

Tables 1 and 2 show the structural properties for pure SM in this and previous works. To the best of our knowledge, there

Table 2. Comparisons of Average Number of Hydrogen Bonds for Various Donor/Acceptor Sites of Pure SM

<table>
<thead>
<tr>
<th>donor/acceptor</th>
<th>atom pair</th>
<th>coordination number</th>
<th>Kehalshahi ( ^{et al.} ) ( ^{38} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>SM–water</td>
<td>OMP or OM10–HW</td>
<td>1.11</td>
<td>1.14</td>
</tr>
<tr>
<td>SM–water</td>
<td>OS7–1F</td>
<td>0.60</td>
<td>0.49</td>
</tr>
<tr>
<td>SM–water</td>
<td>OS11–HW</td>
<td>0.19</td>
<td>0.24</td>
</tr>
<tr>
<td>SM–water</td>
<td>OA30–HW</td>
<td>0.94</td>
<td>0.93</td>
</tr>
<tr>
<td>intermediate</td>
<td>OS17–HS15</td>
<td>0.52</td>
<td>0.30</td>
</tr>
<tr>
<td>intermediate</td>
<td>OA36–HS15</td>
<td>0.18</td>
<td>0.11</td>
</tr>
<tr>
<td>intermediate</td>
<td>HS7–OS11</td>
<td>0.77</td>
<td>0.57</td>
</tr>
</tbody>
</table>

are few experimental studies that report area per lipid for pure SM systems. Further, inherent difficulty in determining accurate area per lipid creates variables and sometimes incompatible results for comparison.\(^{32}\) Hence, we relied primarily on previous simulation studies for comparison. Our computed average area per lipid for pure SM was found to be 53.3 ± 0.4 Å², which is in agreement with simulation values, reported by Niemela et al., of 52 ± 1 Å².\(^{43}\) Simulations on SM bilayers performed by Chiu et al.\(^{31}\) reported area per lipid for 18:0 SM as 53 Å². The simulation performed by these authors used a much larger membrane of 1600 lipids. This larger membrane allowed for more significant effects caused by undulations of the bilayer.

Table 2 shows comparisons of intermolecular hydrogen bonds of SM with previous simulation studies by Kehalshahi et al.\(^{38}\) Interlipid hydrogen bonds between various donor and acceptor sites displayed reasonable agreement with their simulation. Hydrogen bonds between SM and water showed a similar concurrence.

Using the method described by Petracek et al.\(^{44}\) the average volume per lipid for SM at 323 K was calculated to be 1168 ± 12 Å³ and is within error of average volume per lipid calculations of 1180 ± 10 Å³ for 16:0 SM and 1182 Å³ for 18:0 SM from Niemela et al. and Chiu et al., respectively.\(^{34,31}\) Peak-to-peak distance calculated from electron densities reported from Niemela et al. and Chiu et al. were in excellent agreement with our computed values\(^{31,43}\) (Table 1). Further, we note that the average deuterium order parameter for acyl chains was comparable to previous simulation results. Experimental NMR order parameters for 16:0 SM bilayers at 321 K by Mehner et al.\(^{46}\) found an average acyl chain order parameter, \( \left< S_{\text{CD2}} \right> \), of ~0.221, compared with our result of \( \left< S_{\text{CD2}} \right> = 0.246 \pm 0.022 \). Comparison of order parameter profiles from our simulations closely matches profiles from previous simulations.\(^{31,43}\)

There are a limited number of studies on pure Cer bilayers reported in the literature, thereby making validation difficult and limiting the number of conclusions that can be drawn from our data. Differential scanning calorimetry experiments by Shah et al.\(^{33}\) on pure 16:0 Cer bilayers found a broad exothermic transition between 328 and 343 K with no exothermic transition in cooling run.\(^{35}\) This indicated an irreversible transition from a metastable lamellar phase to a stable bilayer. For a fully hydrated bilayer, this transition occurred at 337.2 K. The X-ray diffraction data from Shah et al.\(^{35}\) showed a lamellar repeat distance of 41.8 Å and noted a significant drop in the lamellar repeat distance after the phase transition. Our peak-to-peak distance showed a drop of ~1.3 Å between the temperature of 332 and 340 K. Order parameters also exhibited an overall decrease of ~0.03 for a similar change in temperature and with a corresponding ~1.6 Å increase in area per lipid. Thus, our simulations are successful in reproducing qualitative behavior of a pure Cer bilayer over a wide range of temperatures that encompass a phase transition.

Mixing Properties of the Lipids. Several structural measures can be used to ascertain the mixing behavior of the lipids in bilayers. Our analyses include studies of PMV, PMA, electron densities, and order parameters.

Partial Molecular Volume and Area. Determination of partial molecular volume (PMV) is an effective way of ascertaining the equation of state for heterogeneous liquids. Following Greenwood et al.\(^{47}\) the PMV and partial molecular area (PMA) of the bilayers were investigated. According to the standard definition of PMV, the total volume, \( V = N_{\text{mol}} \langle V_m \rangle \), as a
function of Cer concentration, $x$, for the binary mixture can be expressed as a linear relationship:

\[ v'(\text{Cer}, \text{SM}) = x v_{\text{Cer}} + (1 - x) v_{\text{SM}} \]  

(1)

where $v_{\text{Cer}}$ and $v_{\text{SM}}$ are the PMVs for Cer and SM, respectively. Differentiating eq 1 yields a set of linear equations that can be solved to obtain PMVs as

\[ v_{\text{Cer}} = v(x) + (1 - x) \frac{dv(x)}{dx} \]  

(2)

\[ v_{\text{SM}} = v(x) - x \frac{dv(x)}{dx} \]  

(3)

Figure 2 displays lipid volume as a function of $x$, where extraction of the lipid volume from total box volume was performed using the method proposed by Petricek et al.\textsuperscript{6} We note from Figure 2 that, for most cases, the dependence of the lipid volume on $x$ is a piecewise linear function. Thus, the derivatives in eqs 2 and 3 were calculated by computing the slopes of these fitted functions. The inset graphs of Figure 2 show PMV for SM as a function of $x$. For the values of $x$ between 0.3 and 0.6, $v_{\text{SM}}$ exhibits discontinuities with an overall decrease in PMV. Figure 2d at 358 K did not yield reliable linear or piecewise linear fits. Therefore, $v_{\text{SM}}$ was not computed in this case. Further, as a function of temperature, these discontinuities shift to lower concentrations and are more prominent. Any discontinuity in PMV correlates to volumetric effects in a system due to either chemical reactions or structural rearrangements of the molecules.\textsuperscript{67,68} Classical MD does not incorporate chemical interactions; therefore, structural rearrangements and conformational changes of the lipids must be the cause of any change in PMV.

For two-dimensional liquids such as lipid membranes, area per lipid is considered approximately an extensive quantity similar to the volume.\textsuperscript{69} Hence, one can introduce FMA by essentially replacing volume with area in the volume derivation. Following analysis of Figure 2, we investigated the behavior of PMA (Figure 3). Similar to PMV, the region between 0.1 $\leq$ $x$ $\leq$ 0.6 exhibited a discontinuous decrease in PMA. This discontinuity shifted to higher concentration with increasing temperature. We note that the reduction in PMA was very similar to that of lipid cholesterol mixtures.\textsuperscript{64,68} This behavior implies that the presence of Cer induced structural changes in lipid bilayers comparable with those of Chol and may serve as a surrogate for Chol in this region.

**Deuterium Order Parameters.** Structural changes in lipid membranes are correlated to the degree of disorder in hydrocarbon chains. This disorder is quantified in NMR experiments by measuring deuterium order parameter profiles. The order parameter tensor, $S$, is defined as

\[ S_{ab} = \frac{1}{2} (3 \cos(\theta_a) \cos(\theta_b) - \delta_{ab}) \]

(4)

where $\theta_a$ is the angle made by the $a$th molecular axis with the bilayer normal and $\delta_{ab}$ is the Kronecker delta. In simulations with united atom force fields, the order parameter for saturated and unsaturated carbons, $S_{\text{C,D}}$, can be determined using the following relations: \textsuperscript{69}

\[ S_{\text{C,D}} = \frac{2}{3} S_{xx} + \frac{1}{3} S_{yy} \]  

(5)

\[ S_{\text{C,D}} = \frac{1}{4} S_{xx} + \frac{3}{4} S_{yy} - \frac{\sqrt{3}}{2} S_{xy} \]  

(6)

Figure 4 shows the deuterium chain order parameters averaged over the entire length of the acyl chain (Figure S1 in the Supporting Information). Average sphingosine chain order parameters (Figure S4 in the Supporting Information), calculated from Figure S2 in the Supporting Information, showed similar trends. Careful observation of these figures illustrates that the chain order increases with an increasing
Figure 3. Plots of area per lipid versus $x$. Intersecting linear fits (dashed lines) represent discontinuous changes in slope shown as insets. The slopes of the linear fits correlate to the partial molecular area of SM. Discontinuities in PMA were shifted to higher concentrations for increasing temperatures.

![Graphs showing area per molecule for different temperatures](image)

Figure 4. Average palmitoyl acyl chain order for both Cer and SM molecules as a function of concentration, $x$. Order parameters were averaged over the entire length of the chain. Discontinuity in linear fits (shown in inset) denote possible structural changes. These discontinuities have trends similar to that of PMA. Note that the insets show the derivative of order parameters with respect to Cer concentration, not the actual order parameter values.

![Graphs showing average $S_{22}$ for different temperatures](image)

Concentration of Cer. The rate of change of the order parameter with respect to $x$, plotted as insets, displays discontinuity. This behavior is indicative of Cer’s ordering SM bilayers more effectively at lower concentrations than at higher concentrations. Further, these discontinuities shift toward higher values of $x$ with increasing temperature. This change in order corresponds to similar changes observed in PMA analysis (Figure 3). Discontinuities in the order parameters of sphingosine chains (Figure S4 in the Supporting Information) fluctuate between 0.4 and 0.5 concentrations. Therefore, it can be concluded that the effect of Cer appears to be more prominent on the acyl chain than the sphingosine chain.

Electron Densities. Electron density profiles can also reveal structural properties of a bilayer through analysis of membrane thickness, leaflet symmetry, and lipid packing behavior. These structural analyses can then be correlated to chain order, PMA, and PMV. In experiments, the structure factor of the bilayer is
obtained from X-ray diffraction patterns. The electron density profile is then deduced from the structure factor. In simulation studies, the positions and partial charges of the atoms are known, thereby allowing direct computation of the electron densities. As expected, electron densities show a trough at the center of the hydrocarbon core and prominent peaks corresponding to the phosphate groups (Figure S3 in the Supporting Information). The distance between these peaks ($D_{pp}$) is a measure of bilayer thickness (Figure 5). This figure shows gradual thickening of the bilayer with increasing concentration of Cer. This phenomenon is consistent with the corresponding increase in the order parameters and decrease in PMA. All graphs show abrupt reduction in thickness at intermediate concentrations. This thickness reduction occurs at slightly lower concentration ($x \approx 0.5$ to $x \approx 0.5$) for higher temperatures, 340 and 358 K. The shift in the peak to peak distance appears to have a concentration dependence similar to that of PMV analyses. Since steep reduction in the peak to peak distance cannot be attributed to the smaller volume of Cer alone, the reduction in thickness may be due to possible “lifting” of hydrocarbon chains and/or the configuration of the SM headgroup.

Head Group Orientation. The phosphate group in the SM choline headgroup is, to a significant degree, the most electron dense region of the SM molecule. The position of this phosphate group in the membrane greatly influences the locality of peak electron density. Conformational changes in SM headgroup tilt were quantified by computing the average angle between the bilayer normal and the vector joining the sphingosine C13 backbone carbon to the P8 phosphorus atom (Figure 6A). The figure clearly shows an increase in headgroup tilt with increasing Cer concentration. We note that the tilt angle is largely independent of temperature and alludes to a relationship to Cer concentration consistent with observed decreases in $D_{tth}$.

Head Group-Water Interaction. The change in headgroup orientation may be correlated to a decrease in bound waters surrounding the choline headgroup. Dehydration of SM headgroup may also have an effect on $D_{tth}$ which is a distance between peak electron densities of each leaflet. These electron densities are a superposition of lipid and water densities. Analysis of bound water was performed by calculating radial distribution functions (RDF) of SM head groups with the oxygen atom in water. The average number of bound waters is displayed in Figure 6B. This figure shows an initial hydration of 0.5 waters per headgroup for pure SM followed by a linear increase in the number of bound waters per SM headgroup with respect to Cer concentration. Recent neutron diffraction studies by Ryabova et al. have shown that the incorporation of Cer and Chol into DPPC bilayers increases bilayer thickness by affecting membrane hydration. In addition, the authors noted that Cer, in contrast to Chol, significantly reduces the thickness of the membrane water layer. The effects observed by Ryabova et al. for DPPC bilayers may have correlations similar to SM—Cer mixtures in our systems. The dehydration of SM bilayers with increasing concentrations of Cer is also consistent with observed reductions in hydration of the stratum corneum due to Cer formation by sphingomyelinases.²³

DISCUSSION

Combining analyses of PMV, PMA, average $S_{2CD}$ and $D_{tth}$ we observe emergence of rich and complex mixing behavior in SM—Cer systems. Through enzymatic catalysis of SM, large concentrations of Cer can collect in cellular membranes. Thus, the previously mentioned emergent behavior of these systems may be critical in elucidating mechanisms of some tissues and the biological effects of these molecules and their mixtures. Throughout this work, we evaluated discontinuous changes in observed quantities. Due to finite size and short duration of our simulations, these discontinuities may or may not correspond to thermodynamic phase change; however, in our opinion, these are qualitative indicators of such phenomena.

We observed two distinct correlations based on the six quantities computed in these simulations. Acyl chain $S_{2CD}$ (Figure 4) and PMA (Figure 3) display somewhat equivalent proclivity of discontinuities shifting toward higher values of $x$ with increasing temperature. On the other hand, PMV (Figure 2), headgroup hydration, and $D_{tth}$ (Figure 5) show correlations and are contradistinguished to PMA and acyl chain $S_{2CD}$. These trends are summarized in Figure 7. As noted in previous sections, PMA demonstrates the condensing effect of Cer in SM
bilayers resembling that of Chol. This condensing effect is attenuated at certain concentrations. The loci of these concentrations on a temperature plot demarcate regions of chain orders consistent with separation of liquid disordered and liquid ordered phases. Similar ranges in concentration and temperature have been shown to have enthalpic phase transitions and changes in detergent resistance in experiments performed by Busto et al.\(^2\) (shown as a comparison in Figure 7). The DSC curve in Figure 7 was based on the approximate position of peaks in DSC data by Busto et al. (Figure 7 of Busto et al.\(^2\)). It is important to mention that line fitting has inherent inaccuracy due to noise in the fitted data. The actual points shown in Figure 7 are qualitative indicators of structural change and should not be taken as exact “phase” boundaries.

Hydrogen bonding networks were also analyzed and showed some preferential intermolecular orientations (Table 3). The SM OA36 and H37 hydroxyl group displayed an almost exclusive intermolecular affinity to the OS11 oxygen ester (row 18 of Table 3). The N14 amide group in the sphingosine base evinced a propensity to the acyl chain O17 carbonyl group (row 15 of Table 3). The preferential nature of these intermolecular SM bonds was disrupted by the addition of Cer. In particular, intermolecular SM N14 hydrogen bonds were displaced by an increased affinity for oxygen groups in Cer (rows 7, 8, 17, and 19 of Table 3). Similar bonds in Cer do not show the same structural propensities as that of SM. The OA25 and H26 hydroxyl group showed a slightly increased association with the OS8 acyl chain carbonyl group over the NS sphingosine amide group (rows 12 and 14 of Table 3). The SM N4 quaternary amide group and OM9 and OM10 phosphate oxygens did not play a significant role in interlipid hydrogen bonding. This is contrary to the OS2 headgroup of Cer, which did display significant intermolecular hydrogen bonds with other Cer molecules. At concentrations higher than x = 0.4, the Cer NS amine group showed a substantial increase in affinity for the Cer OS8 carbonyl group and a smaller increase for the Cer OA25 and H26 hydroxyl group (rows 12 and 13 of Table 3). Hydrogen bonds between Cer and SM showed a large degree of fluctuation but did not display any considerable dependence on concentration. Hydrogen bonds for all of our systems exhibited a general insensitivity to temperature. Thus, analysis showed some preferential configuration for SM molecules due to hydrogen bonds between the hydroxyl group and ester group and between the amide group and carbonyl group. These hydrogen bonding patterns were disrupted with increasing Cer concentration. On the other hand, the hydroxyl headgroup of Cer played a far more significant role in intermolecular hydrogen bonding. The hydrogen bonding patterns showed a distinct preference for molecules of the same type (i.e., SM with SM and Cer with Cer).

To the best of our knowledge, the seminal works by Busto et al. are the only experimental results published on bilayers with bimodal mixtures of SM and Cer. Unfortunately, their experiments incorporate Cer concentrations of only up to 50%. In our simulations, a gradual increase in Δ\(\text{H}_{\text{m}}\) is seen for lower concentrations, followed by a significant decrease of 68 Å occurring at concentrations of x ≥ 0.5 (Figure 5). The gradual increase in Δ\(\text{H}_{\text{m}}\) is seen for lower concentrations, followed by a significant decrease of 68 Å occurring at concentrations of x ≥ 0.5 (Figure 5). The gradual increase in Δ\(\text{H}_{\text{m}}\) is seen for lower concentrations, followed by a significant decrease of 68 Å occurring at concentrations of x ≥ 0.5 (Figure 5). The gradual increase in Δ\(\text{H}_{\text{m}}\) is seen for lower concentrations, followed by a significant decrease of 68 Å occurring at concentrations of x ≥ 0.5 (Figure 5). The gradual increase in Δ\(\text{H}_{\text{m}}\) is seen for lower concentrations, followed by a significant decrease of 68 Å occurring at concentrations of x ≥ 0.5 (Figure 5).
normal. The increase in the headgroup angle also corresponds to a decrease in bound waters surrounding the SM headgroup. This may indicate that the SM headgroup is collapsing toward the center of the membrane as a result of the dehydration of the SM headgroup. The drop in PMV may be also be caused by the delocalization of the SM headgroup toward the bilayer center. From the comparisons of these quantities, we may be able to infer that for concentrations of $x \geq 0.5$, the effect of SM headgroup dehydration on $D_{2H}$ surpasses the chain ordering effects of Cer. Experimental results for concentrations above 50% Cer are necessary to verify this observed phenomenon.

Another possible mechanism for the marked decrease in $D_{2H}$ was proposed for pure Cer bilayers by Shah et al.35 They observed that Cer formed a metastable bilayer with straightened hydrocarbon chains that transitioned into a more stable state. Further, they conjectured that this observed transition was caused by the “tilting” of the Cer molecule away from the membrane normal axis. The overall order parameters should be uniformly shifted as a result of such “tilting” behavior. Careful review of the order parameter profiles (Figures S1 and S2 in the Supporting Information) did not exhibit any statistically significant shift. Analysis of the angle formed by the vector from the backbone C13 carbon and terminal acyl methyl groups with the bilayer normal also showed no evidence of this proposed mechanism (data not shown).

The apparent discrepancy between the reduction in $D_{2H}$ and no significant reduction in PMV requires careful study of the methods used in computation of lipid volume. We identify three distinct methodologies to compute lipid volumes, namely:

Method 1: The principle method utilized in this study to determine component lipid volumes was a technique proposed by Petracek et al.39 This procedure details and delineates water volume that can be subsequently subtracted from the total volume to leave only the total lipid volume. Partial molecular volumes of the component lipids can then be determined. This method is ideal, in that it does not explicitly assume any details concerning the conformations of the lipids in the bilayer.

Method 2: Combining membrane area calculations with $D_{2H}$ is another common method of procuring volume data. An estimation of the $z$ axis component of the bilayer can be done using $D_{2H}$ to define membrane thickness. Volume can then be defined as $V_z = D_{2H} \times A/2$. This procedure does assume the lipids are in a straightened orientation in the bilayer and that any deviation from the assumed cylindrical form by conformational changes would cause the $z$ component to shift. These conformational changes would also cause the area per lipid to change, as well, thus conserving the total volume.

Method 3: A method similar to method 2 could be devised using the concept of the Gibbs–Luzzati bilayer thickness, $D_p$ to define a transverse description of the bilayer analogous to $D_{2H}$.45 This method defines a region on the $z$ axis where the lipid-atom-to-water ratio is equal to 1. The distance between these points on both leaflets provides an estimation of the bilayer thickness. Volume can then be calculated as $V_z = D_p \times A/2$. Although, like the previous method, this method also assumes cylindrical shape of the molecule, the thickness calculations are not explicitly dependent on specific lipid conformations. The thickness calculations in this case are only a function of the interfacial properties and the hydration structure of the system.

Figure 8 shows computed volumes using these three methods. At lower concentrations of Cer, $x \leq 0.6$, all three methods show a similar linear decrease in volume per molecule in response to replacing SM with Cer. Volume calculations using method 2 display substantial deviation from the other two methods for Cer concentrations higher than $x \approx 0.6$. In method 2, the positions of the phosphate groups in SM primarily determine $D_{2H}$. Any change in headgroup conformation of SM will be reflected in measurement of $D_{2H}$. We observe that the SM headgroup tilts away from bilayer normal at higher Cer concentrations (Figure 6A). Such tilt, in a mixed bilayer, may provide a configuration in which reduction in $D_{2H}$ can be achieved without
significantly changing area per lipid, thus, reducing the volume by method 2. This mechanism closely resembles the umbrella model introduced by Huang et al. as a possible explanation for the observed solubility of Chol in bilayers.

Simulations at 358 K portray large fluctuations and deviations from fitted data at lower temperatures. These discrepancies may be due to poor performance of the SPC/E water model at high temperatures. Comparison of calculated water density with experimental density values showed diminishing ability of the SPC/E water model to accurately reproduce water densities at relatively higher temperatures (Figure S5 in the Supporting Information). Studies on the thermodynamic properties of water models also show similar effects from SPC/E water at temperatures near boiling point. This leads to the conclusion that possible error in our analyses was due in large part to breakdown of the SPC/E water model at higher simulation temperatures. However, we are confident that our simulation parameters and the SPC/E water model for lower temperatures produce structural properties of our systems with reasonable accuracy.

**SUMMARY**

We have performed molecular dynamics simulations of hydrated 16:0 sphingomyelin bilayers with increasing concentrations of 16:0 ceramide at 333, 332, 340, and 358 K. To the best of our knowledge, these are the only simulations on Cer–SM systems. Through the analysis of PMA, PMV, average chain order parameters, and electron densities, we have identified regions of significant structural changes. These changes coincide with phase changes observed in DSC experiments. Hydrogen bond analysis showed some preferential configuration for SM molecules as a result of hydrogen bonds between specific chemical groups of SM. These observed affinities are disrupted with increasing Cer concentration. Hydrogen bonding patterns also showed a distinct preference for molecules of the same type. Further, we predict significant structural changes in our systems, associated with the dehydration of the SM headgroup, for Cer concentrations above 50%. We also observe that at higher Cer concentrations, SM head groups tilt toward the bilayer center, forming an “umbrella” over Cer under it. Experimental validation of this phenomenon would provide great insight into the structure and function of these systems.

**REFERENCES**

Appendix 2: Evaluation of digestion methods for analysis of trace metals in mammalian tissues and NIST 1577c
Evaluation of digestion methods for analysis of trace metals in mammalian tissues and NIST 1577c

Grace A. Binder, Rainer Metcalf, Zachary Atlas, Kenyon G. Daniel

University of South Florida, Tampa, Chemistry Department & Florida Center for Drug Discovery & Innovation (CDDI), USA

ARTICLE INFO

Keywords:
ICP-OES
Pancreatic cancer
Trace metals
Digestion method
NIST 1577c
ICP-MS

ABSTRACT

Digestion techniques for ICP analysis have been poorly studied for biological samples. This report describes an optimized method for analysis of trace metals that can be used across a variety of sample types. Digestion methods were tested and optimized with the analysis of trace metals in cancerous as compared to normal tissue as the end goal. Anthropological, forensic, oncological and environmental research groups can employ this method reasonably cheaply and safely whilst still being able to compare between laboratories. We examined combined HNO3 and H3PO4 digestion at 175 °C for human, porcine and bovine samples whether they are frozen, fresh or lyophilized powder. Little discrepancy is found between microwave digestion and PFA Teflon pressure vessels. The elements of interest (Cu, Zn, Fe and Ni) yielded consistently higher and more accurate values on standard reference material than samples heated to 75 °C or samples that utilized HNO3 alone. Use of H3PO4 does not improve homogeneity of the sample and lowers precision during ICP analysis. High temperature digestions (> 165 °C) using a combination of HNO3 and H3PO4 as outlined are proposed as a standard technique for all mammalian tissues, specifically, human tissues and yield greater than 300% higher values than samples digested at 75 °C regardless of the acid or acid combinations used. The proposed standardized technique is designed to accurately quantify potential discrepancies in metal loads between cancerous and healthy tissues and applies to numerous tissue studies requiring quick, effective and safe digestions.

Introduction

Trace metals play a vital role in cellular processes as cofactors for enzymes, and in the creation of vascular tissue, cell proliferation and apoptosis [1]. The study of trace metals in normal cellular processes is critical to further our understanding of their role in abnormal situations [2]. Although many trace-metals were examined in this study, four (Cu, Zn, Fe and Ni) were chosen due to their significance in metabolism and, more specifically, disease [3]. Some examples of significant enzymes that utilize trace metals include: matrix metalloproteinase 9 which, upon laceration of the epithelial tissue, utilizes copper to aid in clearance of oxidative damage by macrophages [4]; cytochrome C oxidase of the electron transport chain contains two Cu centers; superoxide dismutase, an important antioxidant, employs Zn or Cu [5]. Deciphering the role of metals in tumorgenesis may assist in better detection, diagnosis and treatment of cancer [6].

A review of the literature over the past 40 years reveals a lack of consistency in sample preparation and digestion techniques of tissue for ICP-MS/ICP-OES (Inductively Coupled Plasma Mass Spectrometry/Optical Emission Spectrometry) analysis. Careful review of prior studies reveals significant differences in digestion protocols across a wide variation of trace metal concentrations reported for similar samples [7–13]. Many of the methods tested did not result in complete digestion of the sample and/or produced wide variation in results. A uniform digestion method is necessary for across laboratory comparison of patient tissue samples in order for accurate and precise information to be gleaned from the trace metal composition in tissues.

Moreover, the aforementioned digestion protocols lead to vastly different measurements of trace metals regardless of the specific instrument. Samples were run on both ICP-MS and ICP-OES concurrently to eliminate disparity based on instrumentation. For the purposes of clarity, only OES data is shown with the exception of human samples.

Background

The main focus of this study is to determine if the variations reported for trace element concentrations in human and animal tissues are due to differences in the tissues themselves or due to sample preparation. A common practice is to employ cross laboratory analysis of digested standard reference material (SRM) similar in matrix to your
samples, but the use of ICP and limited number of certified standard reference materials for a wide range of elements to assess the metal load of mammalian tissues has thus far been limited [14-16]. Published literature has employed numerous heating techniques involving HNO₃, H₂O₂, aqua-regia and HF [10,17,18]. Some procedures suggest that complete digestion can be achieved with HNO₃ alone [7,19,20].

This work seeks to rectify missing information about differences between digestion methods and will show that complete digestions are only achieved using methods similar to digestion II described by Ashoka and colleagues [7], but at increased temperatures. Moreover this study bears to question how much variation in published data on metal loads is due to the sample or the sample preparation technique. As tissue metal research progresses, it is vital to accurately quantify abnormalities in metal loads of pathological tissues [21]. Determination of variability between patients and tissue types is critical and must not be confounded by variations in sample preparation, such as in digestion procedures [9]. Drug discovery, animal biology, forensic science and anthropology all utilize ICP technology for qualitative and quantitative analysis of sample composition making it a bifurcating technology for standardized metalloine analysis.

Method

In this work we specifically focus on Cu, Zn, Ni and Fe and also examine a number of other lesser-studied metals in various tissue types. Early work by Mulay et al. [15] did not show significant differences in the abundance of the metals tested for cancerous and non-cancerous tissues of various types. More recent work concerning mammalian tissue varies widely in digestion protocol, postmortem time and instrument [14,22-24].

Significant improvements in analytical instrumentation allow us to investigate many other metal concentrations at only a few ng/L and lower. Numerous digestion techniques of various foods, sediments, compost as well as animal and human tissues have recently been investigated [7,13,19,20] and generally indicate that a combination of HNO₃ and H₂O₂ heated for 24 h at 75 °C is sufficient to liberate most trace metals (method IV of Ashoka et al. [7]). We tested and modified these above methods for sample preparation and wet tissue digestion but without the assistance of a microwave oven, thereby simplifying it and making it more accessible to a wide variety of laboratories. Further, we utilized five different digestion techniques and evaluated the completeness of each, both visually and chemically, with use of a NIST (National Institute of Science and Technology) standard.

Analytical conditions on the ICP-OES and ICP-MS are summarized in Table 1. Instruments were tuned according to standard methods whereby the ICP-OES was centered and maximized on Mn at 265 nm and the ICP-MS was optimized for the full mass range using a 1 ppb solution of Mg, In and U. The ICP-MS was also optimized to limit oxide interference using CoO/CO and double charges using Ba⁺⁺ /Ba to less than 3%. Series standards were developed from single element stock solutions obtained from either High-Purity Standards (Charlestown, SC, USA) or from Alfa Aesar (Tewksbury, MA, USA) and diluted accordingly to make a linear series of calibration standards. Pig tissues were run along with certified reference material NIST 1577c, a synthetic matrix matched standard, and analytical blanks which were run every 10 samples to ensure both precision and accuracy as well as to account for any drift and assesse carryover between samples. Additional runs included multiple NIST 1577c standards and some runs were entirely comprised of these standards digested in the five methods described.

Digestion of tissues

Organs were freshly harvested from a wild hog/domestic pig hybrid and immediately cooled with dry ice to minimize biochemical degradation in the cell composition from living tissue. The tissue was then stored at ~80 °C until sample preparation.

To maximize surface area and homogenize the sample, we used a crushing technique using liquid nitrogen to completely freeze the sample and then crush it into a fine powder using an agate mortar and pestle. Unfortunately, adipose quickly thawed during the crushing process causing the fats to smear rendering the sample unusable. High temperature ashing techniques were explored and quickly abandoned because so much material was lost due to volatilization in an open vessel that only 1-2 mg maximum was usable from a sample of 100 mg or larger. After consultation with the tissue bank at Moffitt Cancer Center (Tampa, FL, USA) we quickly realized that biospies are typically available in frozen weights of 50 mg or less and therefore sought a suitable alternative. Here we present the most pragmatic methods for wet tissue analysis.

We digested a series of pig tissues in 50 mg samples from pancreas, liver, kidney and lung along with NIST SRM 1577c. For all analyses, following addition of HNO₃ samples were digested for the designated incubation time and were allowed to cool before the addition of the second reagent (either H₂O₂ or H₂SO₄ see Table 1). HNO₃ at high temperature and pressure will effuse from the vessel if improperly cooled, leading to personal safety hazards and loss of material due to volatilization. Finally, after the second incubation period, samples were diluted to 10 mL using highly purified Milli-Q water.

Additionally, since closed instead of open vessels are used as seen in Ref. [22], the amount of material that can be volatilized and lost is reduced maximizing the ability to produce comparative data sets. We opted to cut the sample after minimal thawing from ~80 °C using a ceramic knife and an HDPE non-porous cutting board to minimize metal contamination, similar to the procedures of Rahil-Khaen et al. [14]. The wet weight was recorded and the sample was subsequently placed in a PFA digestion vessel (Savillex corporation, 23 mL PFA pressure vessels for blanks and standards and 5 mL for samples), whereupon reagents were added immediately. The vessels were sealed and placed on a hot plate. Trace metal grade HNO₃ was doubly distilled and then used. ACS certified H₂O₂ was used.

Results

Sample homogeneity was initially checked visually. Samples digested with low heat failed to dissolve all sample material due to residue adhering to the side of the vessel even after decanting. Secondary heating with H₂O₂ after high temperature HNO₃ treatment efficiently oxidized lipids resulting in clear solutions free of precipitates. Furthermore, digestions work best when HNO₃ is added and heated prior to the addition of H₂O₂ thereby limiting loss of sample from evaporation and/or over-pressure which may result in catastrophic failure of the pressure vessel.

NIST 1577c was used as a standard reference material in order to test the aforementioned methods. As outlined previously, the metals analyzed were chosen due to their metabolic importance.

Additional tests incorporating HF, depicted below, adequately dissolved standards but failed to breakdown cancerous breast tissue. These images are notable for the prominent lipid micellation in the form of black precipitates even after addition of HF and heating. HF was therefore avoided due to the volatile and dangerous nature of this acid and other means were sought to digest fatty tissues. It was found that

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Five comparative methods used to digest 50 mg of wet tissue for ICP-OES analysis.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method 1</td>
<td>Combined HNO₃ + H₂O₂ digestion for 12 h at 160 °C</td>
</tr>
<tr>
<td>Method 2</td>
<td>Multiple acid attack utilizing a combination of HNO₃ and H₂SO₄ 160 °C</td>
</tr>
<tr>
<td>Method 3</td>
<td>HNO₃ digestion overnight at 85 °C</td>
</tr>
<tr>
<td>Method 4</td>
<td>H₂O₂ digestion at 175 °C for 4 h followed by addition of H₂O₂ with an additional 2 h heating at 75 °C</td>
</tr>
<tr>
<td>Method 5</td>
<td>same as 1 but with 24 instead of 12 h incubation</td>
</tr>
</tbody>
</table>
sample homogeneity could be achieved through longer incubation periods and higher temperatures without the need for HF.

Evaluation of above methods suggests a combination of HNO₃ and H₂O₂ completely digests the SRM for analysis by ICP technology. Fig. 2 shows that methods 3 and 4 are both inadequate to fully digest samples. The low values observed in method 4 illustrate that, despite initial high temperatures, without significant heating after addition of H₂O₂, precipitation likely sequesters a substantial concentration of the elements tested (Fig. 1).

In an attempt to bridge these methodologies to human tissues, porcine tissues were evaluated to explore the effectiveness of these digestion methods on mammalian tissue. This was done to minimize waste of limited human samples. Percent relative standard deviation is compared across four porcine tissue types to determine the method with the lowest variance (Fig. 3).

The information in Fig. 3 showing the standard deviation for repeated acquisitions of the porcine samples and NIST standard shows that method five has the lowest variance and is therefore the best for

![Graph image](image-url)
generating comparative data sets.

Porcine pancreatic tissue was analyzed in order to determine a window of analysis for four pancreatic cancer biopsies for which we had only a very limited sample (something we expect to be common for most analytical studies on human tissue).

Although the pulverized, dried SRM had much higher element concentrations when compared to wet, whole porcine tissue, digestion method 5 is the best procedure to follow for both sample types. Therefore, method 5 is put forth as a standard to analyze trace metals that is both sensitive and accommodates the degrees of indigestibility exhibited by the various tissue types.

Discussion

The NIST 1577C bovine liver standard is the best quantified material that closely matches our samples with a wide range of well-known trace element concentrations. Additionally, we developed a synthetic matrix matched standard to aid in determination of method precision and instrument drift. Finding a suitable reference material to verify accuracy and creating matrix matched calibration standards is vital to the analytical process to ensure accurate results on human tissues (see discussions by Mulay et al. and Rahil-Khazen et al. [14,15]).

Preliminary data from four pancreatic cancer biopsies with coupled, normal proximal tissue evaluates Cu/Zn ratio based on or digests and it can be seen that in 3 out of 4 cases the ratio showed higher Cu/Zn in tumor tissues (Fig. 4). This preliminary work exemplifies the need to
have accurate high precision data that is comparable across labs and therefore shows the importance of not only having a method that allows for repeatable determination of concentration but also highlights the need for matrix matching standards to ensure accuracy.

Minimizing problems stemming from digestion procedures is one of the critical aims of this study as Moore et al. pointed out that none of the modern ICP methods are completely free from analytical problems [8,10,18,25,26].

HNO₃ is commonly employed in ICP methodology because it has the advantage of not adding additional elements to the plasma that may result in significant matrix effects such as seen with HCl or H₂SO₄ [8]. Matrix-induced spectral overlap problems can occur from ionization potentials, nebulizer flow rates as well as heavy elements or high matrix concentrations [26]. Furthermore, once the plasma atomizes the sample, the addition of hydrogen, oxygen and nitrogen are insignificant due to the atmosphere surrounding the plasma. HNO₃ also has the advantage of producing soluble salts of nearly all metals [25]. However, HNO₃ lacks the ability to completely solubilize fats. Therefore, H₂O₂ was employed for lipid peroxidation.

The use of other acids such as HCl or HClO₄ produces specific mass interferences and can also interfere with complete ionization of the sample [19]. As illustrated by Chamberlain et al., omission of HCI during digestion of fish tissue samples reduced the inorganic molecular interferences for analysis of the carbon-rich samples on the ICP-MS [17]. These methods are properly modified as a way of avoiding the use of expensive instrumentation and dangerous reagents, such as perchloric acid, while still being able to detect low-level trace elements in milligram size samples.

Method 5 was optimized for very small samples to accommodate pancreatic biopsies from tissue banks in 10-50 mg amounts. The use of sealed Teflon pressure vessels minimized material lost in the gaseous phase, though volatilized material may still escape when reagents are added between incubation periods.

Lachas et al. stated that HNO₃ alone cannot solubilize all of the mineral phases present in their coal samples [19]. Additionally, they showed that digestion procedure rather than initial mass of a sample (even in the milligram range) determines the accuracy of the results.

Miscelation in fatty samples causes incomplete extractions of metals from tissue. This was ameliorated in the presence of H₂O₂ through lipid peroxidation. While in most cases, HNO₃ digestion coupled with H₂O₂ at high temperatures yielded the highest counts for element extraction, method 2 utilizing HNO₃-H₂SO₄ produced lower percent error from accepted values for NIST 1577C. It is possible that while method 2 may be better for pulverized, dried tissue; method 5 may more completely extract elements from wet tissue containing higher levels of lipids such as breast tissue. NIST suggests multi acid attack for bovine liver ICP-MS analysis, similar to geologic digestions for whole rock (after Kelley et al., 2003) however digestion method 5 was consistently close to the known value suggesting a more simple technique is just as valid. Assistance of H₂O₂ in completion of digestion is contingent on tissue type because organs, such as the liver, have more adipose and/or fibrotic tissue than others. Although Frame et al. claimed significantly higher blank levels with a matrix including H₂O₂ [8], this was not our experience.

Early experimentation with reagent combinations did not solubilize lipids and often resulted in a residue of fat on the digestion vessel or particulates in the digestate (Fig. 1). Similarly, open vessel digestions resulted in effervescence and loss of material as well as samples with substantial particulates that would clog peristaltic tubing of an instrument and contribute to significant analytical drift.

Digestion method 5 was, overall, best at digesting a variety of tissues across a wide range of elements and creates clear solutions easily dilutable and is highly reproducible. We still caution that this may vary for tissue samples with large amounts of adipose and/or fibrotic tissue. Additional work must also be done to assess the efficacy of this technique for lower atomic masses that are difficult for the ICP-MS to resolve due to mass bias [25].

Conclusions

A standard method of analysis of metals in small tissue samples is vital to obtain comparable inter-laboratory data sets. Differences in the sample preparation can lead to differing results that do not allow adequate determination of the natural variation of metal levels in human tissues. We highly recommend the use of a multi-step high heat digestion technique. The digestion process should be initiated by using HNO₃ alone at temperatures ≥ 175 °C in order to break down all tissue types and to oxidize a significant proportion of carbon. Addition of H₂O₂ must be subsequent to this heating process to eliminate significant volatilization of the sample often resulting in sample material adhering to the walls of the digestion vessel thereby being separated from the digestate. Caution should be taken against chlorinated acids for ICP-MS use. Our results indicate that these chlorinated acids are not needed, neither to break down the tissues further, nor stabilize most elements of interest.

Commercial microwave digestion of organic samples is expensive while perchloric/HF digestion is dangerous. Our method provides a cost effective, safe yet sound alternative to previously published digestion techniques. Although a standard method has never been proposed for digestion of mammalian tissue, it is hoped that it will aid in the advancement of cancer research by quantifying trace metal loads. Future research seeks to apply our preferred method to whole tissue comparison of breast tumors to non-cancerous breast tissue using our preliminary data on metal loads inspired by this work.

Acknowledgements

The authors would like to acknowledge Cooper the pig and Fibonda Williams from for their contribution to this work as well as Michelle Wilde.

References


Appendix 3: Cupriphilic compounds to aid in proteasome inhibition
Cuprophilic compounds to aid in prosesosome inhibition

Sreya Mukherjee, Robert Sparks, Rainer Metcalf, Wesley Brooks, Kenyon Daniel, Wayne C. Guida

Department of Chemistry, University of South Florida, CHE 205, 4202 E. Fowler Avenue, Tampa, FL 33620, USA
Department of Cell Biology, Microbiology and Molecular Biology, University of South Florida, 4202 E. Fowler Ave, Tampa, FL, USA

A R T I C L E   I N F O
Article history:
Received 25 January 2016
Revised 21 April 2016
Accepted 23 April 2016
Available online 25 April 2016

Keywords:
Cancer
Prosesosome
Cuprophilic compounds
Virtual screening

A B S T R A C T
It has been found that tumor cells and tissues, compared to normal cells, have higher levels of copper and possibly other metal ions. This presents a potential vulnerability of tumor cells that can serve as a physiological difference between cancer cells and normal cells and allows design of compounds that selectively target tumor cells while sparing normal cells. Recently we have identified compounds that have potential to inhibit the prosesosome in tumor cells and induce cell death by mobilizing endogenous tumor copper resulting in in cellulo activation of the compound. These compounds hence act as prodrugs, becoming active drugs in tumor cells with high copper content but remaining essentially inactive in normal cells, thereby greatly reducing adverse effects in patients. Such use would be of significant benefit in early detection and treatment of cancers, in particular, aggressive cancers such as pancreatic cancer which is usually not detected until it has reached an advanced stage. Six compounds were identified following virtual screening of the NCI Diversity Set with our prosesosome computer model followed by confirmation with a biochemical assay that showed significant inhibition of the prosesosome by the compounds in the presence of copper ions. In a dose response assay, NSC 37408 (6- dihydroxy-1- benzofuran-3-one), our best compound, exhibited an IC50 of 3 μM in the presence of 100 nM copper.

© 2016 Published by Elsevier Ltd.

In cancer, therapy can be as debilitating as the disease itself considering that current treatments are often accompanied by severe toxicities in patients. These toxicities prompt continuing investigation into new therapies with reduced, or preferably, no toxic effects. Whereas treating cancer cells without toxicity in normal cells is the ultimate goal of drug discovery, this task has met with limited success due in large part to the difficulty of finding therapeutics that can distinguish cancer cells from normal cells.

Copper, which has the ability to adopt both oxidized (Cu2+) and reduced (Cu1+) states, is an essential trace element for various metabolic processes in living organisms. There are several enzymes that use copper for processes necessary for carcinogene-
sis such as extracellular matrix degradation, endothelial cell proliferation, and migration mediated by integrins. Due to its role in important physiologic processes, including metabolism, the concentration of copper in organisms is tightly regulated. Copper is an element that plays an essential role in tumor development, angiogenesis, and metastasis. Experimental evidence exists that shows tumor tissues possess both elevated copper and altered copper/zinc ratios in a stage dependent manner across multiple types of carcinomas. However, these studies of the disposition of trace metals within cancer patients comparing normal tissues to cancerous tissues have been limited. In most of the studies focusing on copper, zinc, iron, and selenium, it has been observed that the copper concentrations, as opposed to other metals, were almost always found to be elevated compared to age matched samples from normal tissue. For example, elevated serum copper levels in cancer patients have been reported in a wide variety of tumors in the following tissues: breast, cervical, ovarian, lung, prostate, and stomach. The biomolecular target of the present study, the prosesosome, contains a 700 kDa barrel-shaped core particle formed by four axially stacked heptameric rings. These rings include the β subunits that possess six proteases whose active sites face an interior lumen where proteolysis occurs. These active sites further possess a cat-
alytic region (S1) which includes the catalytic THR1 residue and recognition region (S3). The 20S prosesosome, which is the proteolytic core of the 20S prosesosome complex, contains multiple peptidase activities including the chymotrypsin-like (CTL), trypsin-like (TL), and peptidylglutamyl peptide hydrolyzing (PGPH) or caspase-like activities. The prosesosome is involved in oncogenic events such as up-regulation of cellular proliferation, angiogenesis, down regulation of apoptosis, and drug resistance. Pursuing prosesosome inhibition is a clinically validated therapeutic strategy. For example, prosesosome inhibition is believed to result in buildup of abnormal or damaged proteins that triggers apoptosis. Figure 1 depicts the 20S prosesosome containing β subunits 4, 5, and 6 which...
are important in mediating the CTL activity of the proteasome. Copper is placed here in cyan based on quantum calculations discussed later. Note that copper is not required for proteasome activity.

In prior studies, we have shown that addition of copper complexation agents (independent of the salt form) to copper, had similar potency to inorganic copper salts in purified protein assays.** However, these compounds displayed substantially increased potency on human leukemia cancer cell lines. Thus, treatment of Jurkat T-cells with an 8-hydroxyquinoline copper mixture induced loss of viability in cell cultures. This effect was not observed upon treatment of Jurkat T-cells with 8-hydroxyquinoline alone or copper alone.

We previously also performed experiments on immobilized MCF-10A, premalignant MCF10AT1.K1c12, malignant MCF10DCIS.com and normal MDA-MB-231 breast cell lines, which were treated with copper complexing compounds cloquinoil or pyrrolidine dithiocarbamate, in the presence and absence of copper, followed by measurement of cell survival rate. CTL activity of proteasome inhibition and apoptosis were also determined by levels of the ubiquitinated proteins in protein extracts of the treated cells and showed that cloquinoil or pyrrolidine dithiocarbamate become active proteasome inhibitors and breast cancer cell killers in the presence of copper. These prior literature and experimental facts have brought to light the possibility of targeting tumor growth, angiogenesis, and metastasis with cupriche compounds.** In an attempt to realize this goal, our lab is using small molecules with a propensity to bind copper, cost effective synthesis, and a high potential for rational synthesis of analogues. With elevated copper levels as a marker, we hypothesize that our compounds can be used specifically against tumor cells with minimal impact on normal cells and may alleviate toxicities seen in current chemotherapeutic treatments.

Previous studies show that in cells assembled copper-activated proteasome inhibitors have apoptosis-inducing effects on a wide array of solid tumors and no measurable effect on normal cells.** Yet, the field of copper-activated proteasome inhibitors has stalled due to lack of therapeutically suitable compounds. Only a very small number of organic scaffolds have been studied with respect to complexation with copper for proteasome inhibition in cancer cells, including: pyrrolidine dithiocarbamate, 8-hydroxyquinoline (8-HQ), cloquinol (CQ), and dusulfiram. Prior studies have shown that these compounds have differential effects in immortalized, pre-malignant, and malignant breast cancer cells.**

In the present study, our laboratory has utilized computational screening followed by experimental testing in an in vitro biochemical assay to discover a new set of compounds that, according to the hypothesized mode of action, should be selective for induction of apoptosis in tumor cells via proteasome inhibition. The compounds identified presumably could act as pro-drugs, becoming an active drug in the presence of copper in tumor cells, and then inducing apoptosis by proteasome inhibition.

Virtual screening was performed on compounds contained within the NCI Diversity Set 3 against the 20S proteasome model. Schrödinger’s Maestro 9.3.5 was used as the primary graphical user interface for molecule structure preparation and Schrödinger applications were used for analysis. Quantum mechanical refinement of copper interactions with the THR1 in the active site using Q-site** and Jaguar** with B3LYP/LACVP allowed for placement of copper and the assignment of partial charges on THR1 and the copper ion. The virtual screening method employed the modified yeast 20S proteasome crystal structure derived from PDB ID: 1R1.** Ligands from the NCI Diversity Set 3 were prepared with LigPrep** and metal binding sites were added for generation of appropriate ligand states to interact with the copper ion. The standard precision (SP) setting in Glide was used for docking to incorporate metal binding sites.** Out of 1597 compounds, 62 were selected by the virtual screening method, which were then tested at 10 μM in the presence of 1 μM copper. A Beckman Coulter Biomek FX® Lab Automation Workstation was used for the automated assay runs. The assay was performed in 384-well black Nunc plates. The process involved the addition of 2 μl compound in DMSO added to 28 μl buffer (50 mM Tris; pH 7.6, 37 °C), with 20 μM, 10 μM 20S proteasome, and 10 μM of succ-leu-leu-val-tyr-AMC substrate and the rate of substrate cleavage/20S proteasome activity was determined. The assay in the presence of copper chloride solution involved addition of 3 μl of the copper solution and decrease in the amount of buffer solution to 25 μl. To allow chelation of copper to the compound, the plate was allowed to sit for 40 min with gentle shaking. The overall volume per well was kept constant at 50 μl. The compounds and substrate were initially dissolved in 100% DMSO, but the final concentration of DMSO per well plate was brought down to below 2% following subsequent dilutions. Plates were incubated at 37 °C for 2 h. A Perkin Elmer 2102 multi-label plate reader was used for fluorescence measurement. The plates were read using 340 nm excitation and 460 nm emission filters. All liquid transfers to the plates were performed using the Biomek Workstation. Six lead compounds, depicted in Figure 2, exhibited greater than 80% proteasome inhibition in the presence of 1 μM copper. Disulfiram, a well-known copper chelating compound, and copper alone were used as controls in the experiment. For dose-response curves, the concentrations of copper chloride and compound were varied to find the greatest percent inhibition and results have been reported herein.

To understand whether other metals were interacting with the compound, a solution of metals such as nickel, calcium, iron, zinc and copper chloride were taken and tested in the same way as mentioned above with the compound. The inhibition of proteasome was to the same order as that for copper chloride for the same conditions suggesting that this is a copper specific phenomenon.

A series of novel copper-activated proteasome inhibitors was identified from NCI Diversity Set 3 that exhibited greater than 80% inhibition of the 20S proteasome when assayed at 10 μM in the presence of 1 μM copper. In a dose-response assay, NSC 37408 gave the best results with an IC50 of 3 μM in presence of 100 nM copper.

Figure 3 depicts the dose-response curves for proteasome inhibition with copper alone, compound alone, and compound with 100 nM copper. This shows a dramatic improvement with the compound in the presence of 100 nM copper, while copper alone or compound alone at that concentration exhibits nominal activity.
Figure 2. Compounds that show proteasome inhibition.

Figure 3. The purified protein dose response curve for NSC 34708 showing copper alone (blue), compound alone (black), and compound with copper (red) depicts a dramatic improvement of percent inhibition by the compound in the presence of submicromolar copper. Copper alone at that concentration exhibits little to no observable inhibitory activity (each experiment was done in triplicate, relative error <1% for all 3 experiments).

Disulfiram, a known copper chelating compound, was used as a positive control which exhibited 92% inhibition of CTL activity at 10 μM when assayed in the presence of 1 μM copper, whereas NSC 34708, showed 80% inhibition at 10 μM with 1 μM copper. This compound is being pursued because it is amenable to lead optimization and it possesses a unique chemical scaffold compared to previously used cuprophilic proteasome inhibitors. Many of the top-ranking compounds, as depicted in Figure 3, have phenolic hydroxyl groups and carboxylic acids moieties which have the propensity for copper binding. The coordination of these functional groups to copper is well understood in the chemical literature. As demonstrated by the assay, this class of cuprophilic compounds has the ability to bind copper and enhance proteasome inhibition, even in the submicromolar range. Figures 4 and 5 demonstrate a possible mode of action and binding sites of these compounds in the active site of the proteasome. As per our model and docking methods used for NSC 37408, the C7 phenolic hydroxyl group (7-OH)

Figure 4. Glide SP-docking pose of NSC 37408 (shown as green tubes) bound to copper (shown in cyan) in the active site. The C7 subunit of the proteasome is shown in magenta.

The hypothesis considered for building the proteasome model is that Cu (II) coordinates to the NH₂ and OH groups of THRI in the active site, followed by binding of our compounds to this complex. Assuming coordination of Cu (II) with THRI, quantum mechanical modeling of Cu (II) interactions with THRI was done using Q-Site, which assisted in the placement of copper into the j5 active site and the assignment of appropriate partial charges on THRI and copper. Virtual screening conducted on the ligands from NCI Diversity Set 3 allowed the selection of top-ranking compounds that were then assayed for inhibition of the CTL activity against purified 20S proteasome. Each compound was tested in the presence and absence of copper chloride; 6 were found to inhibit the proteasome with greater inhibition at 10 μM when assayed in the presence of 100 nM copper and no or sometimes minimal inhibition was observed without copper. As previously known, copper alone can serve as a weak inhibitor of the proteasome under our assay conditions and, therefore, was included as a control in the experiment.
Figure 5. NSC 37408 (shown in green) binds to copper (shown in cyan) into a well-defined pocket between the residues THR21, THR1 and GLY47 in the θ5 subunit of the catalytic site of the proteasome.

and the other moiety are expected to coordinate with copper. pKₐ calculations with Jaguar suggest a pKₐ of 3.9 for the 7-OH substituent, consistent with a vinylogous carboxylic acid as shown in Figure 5. Additionally, the θ5 subunit has a well-defined binding pocket formed between THR1, THR21 and GLY47 which could potentially harbor the copper complex.

In conclusion, rational drug discovery methods resulted in the successful identification of compounds that inhibit the CTL activity of the proteasome following their coordination to copper as suggested by the computational and biochemical assay results. The computational methods helped remove 90% of the compounds (only 62 compounds out of 1597 were tested) and the biochemical assay was able to detect decoys and further remove 90% of false positives from the virtual screen (6 true positives out of 62). The compound pursued (NSC 37408) exhibited a 3 μM IC₅₀ against the proteasome in the presence of 100 nM copper (with nominal inhibition in the absence of copper), and further SAR studies on this compound are planned to showcase its amenability to design of analogues with better in vitro and in vivo properties. Determination of inhibition of other subunits of the proteasome will be pursued in due course.

Acknowledgements

This work was supported by a Proposal Enhancement Grant from the University of South Florida. We thank Mr. Nan Sun of the Moffitt Cancer Center at the University of South Florida, Tampa for his assistance with assay development.

References and notes

Appendix 4: Fair Use Worksheets

Appendix 4.1: Ligand-Mediated Protein Degradation Reveals Functional Conservation Among Sequence Variants of the CUL4-type E3 Ligase Substrate Receptor Cereblon
INSTRUCTIONS

Check all boxes that apply, and keep a copy of this form for your records. If you have questions, please contact the USF General Counsel or your USF Tampa Library Copyright Librarian.

Name: Rainer Metcalf  Date: 11/08/19

Class or Project: Dissertation

Title of Copyrighted Work:

<table>
<thead>
<tr>
<th>Likely Supports Fair Use</th>
<th>Likely Does Not Support Fair Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>☐ Educational</td>
<td>☐ Commercial</td>
</tr>
<tr>
<td>☐ Teaching (including multiple copies for classroom use)</td>
<td>☐ Entertainment</td>
</tr>
<tr>
<td>☐ Research or Scholarship</td>
<td>☐ Bad-faith behavior</td>
</tr>
<tr>
<td>☐ Criticism, Parody, News Reporting or Comment</td>
<td>☐ Denying credit to original author</td>
</tr>
<tr>
<td>☐ Transformative Use (your new work relies on and adds new expression, meaning, or message to the original work)</td>
<td>☐ Non-transformative or exact copy</td>
</tr>
<tr>
<td>☐ Restricted Access (to students or other appropriate group)</td>
<td>☐ Made accessible on Web or to public</td>
</tr>
<tr>
<td>☐ Nonprofit</td>
<td>☐ Profit-generating use</td>
</tr>
</tbody>
</table>

Overall, the purpose and character of your use ☑ supports fair use or ☐ does not support fair use.

<table>
<thead>
<tr>
<th>Likely Supports Fair Use</th>
<th>Likely Does Not Support Fair Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>☐ Factual or nonfiction</td>
<td>☐ Creative or fiction</td>
</tr>
<tr>
<td>☐ Important to favored educational objectives</td>
<td>☐ Consumable (workbooks, tests)</td>
</tr>
<tr>
<td>☐ Published work</td>
<td>☐ Unpublished</td>
</tr>
</tbody>
</table>

Overall, the nature of the copyrighted material ☑ supports fair use or ☐ does not support fair use.

<table>
<thead>
<tr>
<th>Likely Supports Fair Use</th>
<th>Likely Does Not Support Fair Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>☐ Small amount (using only the amount necessary to accomplish the purpose)</td>
<td>☐ Large portion or whole work</td>
</tr>
<tr>
<td>☐ Amount is important to favored socially beneficial objective (i.e. educational objectives)</td>
<td>☐ Portion used is qualitatively substantial (i.e. it is the ‘heart of the work’)</td>
</tr>
<tr>
<td>☐ Lower quality from original (ex. Lower resolution or bit rate photos, video, and audio)</td>
<td>☐ Similar or exact quality of original work</td>
</tr>
</tbody>
</table>

LeEtta Schmidt, lmschmidt@usf.edu and Drew Smith dsmith@usf.edu
Reviewed by USF General Counsel 08/11/2015
Overall, the amount and substantiality of material used in relation to the whole supports fair use or does not support fair use.

**EFFECT ON THE MARKET FOR ORIGINAL**

<table>
<thead>
<tr>
<th>Likely Supports Fair Use</th>
<th>Likely Does Not Support Fair Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>☐ No significant effect on the market or potential market for the original</td>
<td>☐ Replaces sale of copyrighted work</td>
</tr>
<tr>
<td>☐ No similar product marketed by the copyright holder</td>
<td>☐ Significantly impairs market or potential market for the work</td>
</tr>
<tr>
<td>☐ You own a lawfully acquired copy of the material</td>
<td>☐ Numerous copies or repeated, long-term use</td>
</tr>
<tr>
<td>☐ The copyright holder is unidentifiable</td>
<td>☐ Made accessible on Web or to public</td>
</tr>
<tr>
<td>☐ Lack of licensing mechanism for the material</td>
<td>☐ Affordable and reasonably available permissions or licensing</td>
</tr>
</tbody>
</table>

Overall, the effect on the market for the original supports fair use or does not support fair use.

**CONCLUSION**

The combined purpose and character of the use, nature of the copyrighted material, amount and substantiality of material used in relation to the whole and the effect on the market for the original likely supports fair use or likely does not support fair use.

*Note: Should your use of copyrighted material not support fair use, you may still be able to locate and request permissions from the copyright holder. For help on this, please feel free to contact your Copyright Librarian.*

This worksheet has been adapted from:

- Cornell University’s Checklist for Conducting A Fair use Analysis Before Using Copyrighted Materials:  
  [https://copyright.cornell.edu/policies/docs/Fair_Use_Checklist.pdf](https://copyright.cornell.edu/policies/docs/Fair_Use_Checklist.pdf)
- Smith, Kevin; Macklin, Lisa A.; Gilliland, Anne. A Framework for Analyzing any Copyright Problem. Retrieved from:  
  [https://d396quza40or.cloudfront.net/cfei/Reading%20Docs/A%20Framework%20for%20Analyzing%20any%20Copyright%20Problem.pdf](https://d396quza40or.cloudfront.net/cfei/Reading%20Docs/A%20Framework%20for%20Analyzing%20any%20Copyright%20Problem.pdf)

LeEtta Schmidt, [lmschmidt@usf.edu](mailto:lmschmidt@usf.edu) and Drew Smith [dsmith@usf.edu](mailto:dsmith@usf.edu)

Reviewed by [USF General Counsel](mailto:USF-General-Counsel@usf.edu) 08/11/2015
Appendix 4.2: Mixing Properties of Sphingomyelin Ceramide Bilayers: A Simulation Study
**INSTRUCTIONS**

Check all boxes that apply, and keep a copy of this form for your records. If you have questions, please contact the USF General Counsel or your USF Tampa Library Copyright Librarian.

**Name:** Rainer Metcalf  
**Date:** 11/08/19

**Class or Project:** Dissertation  
**Title of Copyrighted Work:** Mixing Properties of Sphingomyelin Ceramide Bilayers: A Simulation Study

### PURPOSE AND CHARACTER OF THE USE

<table>
<thead>
<tr>
<th>Likely Supports Fair Use</th>
<th>Likely Does Not Support Fair Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>☐ Educational</td>
<td>☐ Commercial</td>
</tr>
<tr>
<td>☐ Teaching (including multiple copies for classroom use)</td>
<td>☐ Entertainment</td>
</tr>
<tr>
<td>☐ Research or Scholarship</td>
<td>☐ Bad-faith behavior</td>
</tr>
<tr>
<td>☐ Criticism, Parody, News Reporting or Comment</td>
<td>☐ Denying credit to original author</td>
</tr>
<tr>
<td>☐ Transformative Use (your new work relies on and adds new expression, meaning, or message to the original work)</td>
<td>☐ Non-transformative or exact copy</td>
</tr>
<tr>
<td>☐ Restricted Access (to students or other appropriate group)</td>
<td>☐ Made accessible on Web or to public</td>
</tr>
<tr>
<td>☐ Nonprofit</td>
<td>☐ Profit-generating use</td>
</tr>
</tbody>
</table>

Overall, the purpose and character of your use supports fair use or does not support fair use.

### NATURE OF THE COPYRIGHTED MATERIAL

<table>
<thead>
<tr>
<th>Likely Supports Fair Use</th>
<th>Likely Does Not Support Fair Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>☐ Factual or nonfiction</td>
<td>☐ Creative or fiction</td>
</tr>
<tr>
<td>☐ Important to favored educational objectives</td>
<td>☐ Consumable (workbooks, tests)</td>
</tr>
<tr>
<td>☐ Published work</td>
<td>☐ Unpublished</td>
</tr>
</tbody>
</table>

Overall, the nature of the copyrighted material supports fair use or does not support fair use.

### AMOUNT AND SUBSTANTIALITY OF MATERIAL USED IN RELATION TO WHOLE

<table>
<thead>
<tr>
<th>Likely Supports Fair Use</th>
<th>Likely Does Not Support Fair Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>☐ Small amount (using only the amount necessary to accomplish the purpose)</td>
<td>☐ Large portion or whole work</td>
</tr>
<tr>
<td>☐ Amount is important to favored socially beneficial objective (i.e. educational objectives)</td>
<td>☐ Portion used is qualitatively substantial (i.e. it is the ‘heart of the work’)</td>
</tr>
<tr>
<td>☐ Lower quality from original (ex. Lower resolution or bitrate photos, video, and audio)</td>
<td>☐ Similar or exact quality of original work</td>
</tr>
</tbody>
</table>

LeEtta Schmidt, [lmschmidt@usf.edu](mailto:lmschmidt@usf.edu) and Drew Smith [dsmith@usf.edu](mailto:dsmith@usf.edu)  
Reviewed by [USF General Counsel](mailto:USF.Genera) 08/11/2015
Overall, the amount and substantiality of material used in relation to the whole supports fair use or does not support fair use.

**EFFECT ON THE MARKET FOR ORIGINAL**

<table>
<thead>
<tr>
<th>Likely Supports Fair Use</th>
<th>Likely Does Not Support Fair Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>❑ No significant effect on the market or potential market for the original</td>
<td>❑ Replaces sale of copyrighted work</td>
</tr>
<tr>
<td>❑ No similar product marketed by the copyright holder</td>
<td>❑ Significantly impairs market or potential market for the work</td>
</tr>
<tr>
<td>❑ You own a lawfully acquired copy of the material</td>
<td>❑ Numerous copies or repeated, long-term use</td>
</tr>
<tr>
<td>❑ The copyright holder is unidentified</td>
<td>❑ Made accessible on Web or to public</td>
</tr>
<tr>
<td>❑ Lack of licensing mechanism for the material</td>
<td>❑ Affordable and reasonably available permissions or licensing</td>
</tr>
</tbody>
</table>

Overall, the effect on the market for the original supports fair use or does not support fair use.

**CONCLUSION**

The combined purpose and character of the use, nature of the copyrighted material, amount and substantiality of material used in relation to the whole and the effect on the market for the original supports fair use or does not support fair use.

*Note: Should your use of copyrighted material not support fair use, you may still be able to locate and request permissions from the copyright holder. For help on this, please feel free to contact your Copyright Librarian.*

This worksheet has been adapted from:

Cornell University’s Checklist for Conducting A Fair use Analysis Before Using Copyrighted Materials: [https://copyright.cornell.edu/policies/docs/Fair_Use_Chitelst.pdf](https://copyright.cornell.edu/policies/docs/Fair_Use_Chitelst.pdf)


LeEtta Schmidt, lmschmidt@usf.edu and Drew Smith dsmith@usf.edu

Reviewed by USF General Counsel 08/11/2015
Appendix 4.3: Evaluation of digestion methods for analysis of trace metals in mammalian tissues and NIST 1577c
INSTRUCTIONS

Check all boxes that apply, and keep a copy of this form for your records. If you have questions, please contact the USF General Counsel or your USF Tampa Library Copyright Librarian.

Name: Rainer Metcalf          Date: 11/08/19

Class or Project: Dissertation

Title of Copyrighted Work: Evaluation of digestion methods for analysis of trace metals in mammalian tissues and NIST 1677c

PURPOSE AND CHARACTER OF THE USE

<table>
<thead>
<tr>
<th>Likely Supports Fair Use</th>
<th>Likely Does Not Support Fair Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>Educational</td>
<td>Commercial</td>
</tr>
<tr>
<td>Teaching (including multiple copies for classroom use)</td>
<td>Entertainment</td>
</tr>
<tr>
<td>Research or Scholarship</td>
<td>Bad-faith behavior</td>
</tr>
<tr>
<td>Criticism, Parody, News Reporting or Comment</td>
<td>Denying credit to original author</td>
</tr>
<tr>
<td>Transformative Use (your new work relies on and adds new expression, meaning, or message to the original work)</td>
<td>Non-transformative or exact copy</td>
</tr>
<tr>
<td>Restricted Access (to students or other appropriate group)</td>
<td>Made accessible on Web or to public</td>
</tr>
<tr>
<td>Nonprofit</td>
<td>Profit-generating use</td>
</tr>
</tbody>
</table>

Overall, the purpose and character of your use: □ supports fair use or □ does not support fair use.

NATURE OF THE COPYRIGHTED MATERIAL

<table>
<thead>
<tr>
<th>Likely Supports Fair Use</th>
<th>Likely Does Not Support Fair Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factual or nonfiction</td>
<td>Creative or fiction</td>
</tr>
<tr>
<td>Important to favored educational objectives</td>
<td>Consumable (workbooks, tests)</td>
</tr>
<tr>
<td>Published work</td>
<td>Unpublished</td>
</tr>
</tbody>
</table>

Overall, the nature of the copyrighted material: □ supports fair use or □ does not support fair use.

AMOUNT AND SUBSTANTIALITY OF MATERIAL USED IN RELATION TO WHOLE

<table>
<thead>
<tr>
<th>Likely Supports Fair Use</th>
<th>Likely Does Not Support Fair Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small amount (using only the amount necessary to accomplish the purpose)</td>
<td>Large portion or whole work</td>
</tr>
<tr>
<td>Amount is important to favored socially beneficial objective (i.e. educational objectives)</td>
<td>Portion used is qualitatively substantial (i.e. it is the ‘heart of the work’)</td>
</tr>
<tr>
<td>Lower quality from original (ex. Lower resolution or bitrate photos, video, and audio)</td>
<td>Similar or exact quality of original work</td>
</tr>
</tbody>
</table>

LeEtta Schmidt, lmschmidt@usf.edu and Drew Smith dsmith@usf.edu
Reviewed by USF General Counsel 08/11/2015
Overall, the amount and substantiability of material used in relation to the whole supports fair use or does not support fair use.

**EFFECT ON THE MARKET FOR ORIGINAL**

<table>
<thead>
<tr>
<th>Likely Supports Fair Use</th>
<th>Likely Does Not Support Fair Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>• No significant effect on the market or potential market for the original</td>
<td>• Replaces sale of copyrighted work</td>
</tr>
<tr>
<td>• No similar product marketed by the copyright holder</td>
<td>• Significantly impairs market or potential market for the work</td>
</tr>
<tr>
<td>• You own a lawfully acquired copy of the material</td>
<td>• Numerous copies or repeated, long-term use</td>
</tr>
<tr>
<td>• The copyright holder is unidentifiable</td>
<td>• Made accessible on Web or to public</td>
</tr>
<tr>
<td>• Lack of licensing mechanism for the material</td>
<td>• Affordable and reasonably available permissions or licensing</td>
</tr>
</tbody>
</table>

Overall, the effect on the market for the original supports fair use or does not support fair use.

**CONCLUSION**

The combined purpose and character of the use, nature of the copyrighted material, amount and substantiability of material used in relation to the whole and the effect on the market for the original supports fair use or does not support fair use.

*Note: Should your use of copyrighted material not support fair use, you may still be able to locate and request permissions from the copyright holder. For help on this, please feel free to contact your Copyright Librarian.*

This worksheet has been adapted from:

Cornell University’s Checklist for Conducting A Fair use Analysis Before Using Copyrighted Materials: [https://copyright.cornell.edu/policies/docs/Fair_Use_Checklist.pdf](https://copyright.cornell.edu/policies/docs/Fair_Use_Checklist.pdf)


Smith, Kevin; Macklin, Lisa A.; Gilliland, Anne. A Framework for Analyzing any Copyright Problem. Retrieved from: [https://d3996qasz40orc.cloudfront.net/cfel/Reading%20Docs/A%20Framework%20for%20Analyzing%20Any%20Copyright%20Problem.pdf](https://d3996qasz40orc.cloudfront.net/cfel/Reading%20Docs/A%20Framework%20for%20Analyzing%20Any%20Copyright%20Problem.pdf)

LeEtta Schmidt, [rnschmidt@usf.edu](mailto:rnschmidt@usf.edu) and Drew Smith [dsmith@usf.edu](mailto:dsmith@usf.edu)

Reviewed by [USF General Counsel](mailto:USFGeneralCounsel@usf.edu) 08/11/2015
Appendix 4.4: Cupriphilic compounds to aid in proteasome inhibition
INSTRUCTIONS

Check all boxes that apply, and keep a copy of this form for your records. If you have questions, please contact the USF General Counsel or your USF Tampa Library Copyright Librarian.

Name: Rainer Metcalf Date: 11/08/19

Class or Project: Dissertation

Title of Copyrighted Work: Cuprophilic compounds to aid in proteasome inhibition

PURPOSE AND CHARACTER OF THE USE

<table>
<thead>
<tr>
<th>Likely Supports Fair Use</th>
<th>Likely Does Not Support Fair Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>■ Educational</td>
<td>□ Commercial</td>
</tr>
<tr>
<td>■ Teaching (including multiple copies for classroom use)</td>
<td>□ Entertainment</td>
</tr>
<tr>
<td>■ Research or Scholarship</td>
<td>□ Bad-faith behavior</td>
</tr>
<tr>
<td>■ Criticism, Parody, News Reporting or Comment</td>
<td>□ Denying credit to original author</td>
</tr>
<tr>
<td>■ Transformative Use (your new work relies on and adds new expression, meaning, or message to the original work)</td>
<td>□ Non-transformative or exact copy</td>
</tr>
<tr>
<td>■ Restricted Access (to students or other appropriate group)</td>
<td>■ Made accessible on Web or to public</td>
</tr>
<tr>
<td>■ Nonprofit</td>
<td>□ Profit-generating use</td>
</tr>
</tbody>
</table>

Overall, the purpose and character of your use □ supports fair use or ■ does not support fair use.

NATURE OF THE COPYRIGHTED MATERIAL

<table>
<thead>
<tr>
<th>Likely Supports Fair Use</th>
<th>Likely Does Not Support Fair Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>■ Factual or nonfiction</td>
<td>□ Creative or fiction</td>
</tr>
<tr>
<td>■ Important to favored educational objectives</td>
<td>□ Consumable (workbooks, tests)</td>
</tr>
<tr>
<td>■ Published work</td>
<td>■ Unpublished</td>
</tr>
</tbody>
</table>

Overall, the nature of the copyrighted material ■ supports fair use or □ does not support fair use.

AMOUNT AND SUBSTANTIABILITY OF MATERIAL USED IN RELATION TO WHOLE

<table>
<thead>
<tr>
<th>Likely Supports Fair Use</th>
<th>Likely Does Not Support Fair Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>■ Small amount (using only the amount necessary to accomplish the purpose)</td>
<td>■ Large portion or whole work</td>
</tr>
<tr>
<td>■ Amount is important to favored socially beneficial objective (i.e. educational objectives)</td>
<td>■ Portion used is qualitatively substantial (i.e. it is the ‘heart of the work’)</td>
</tr>
<tr>
<td>■ Lower quality from original (ex. Lower resolution or bitrate photos, video, and audio)</td>
<td>■ Similar or exact quality of original work</td>
</tr>
</tbody>
</table>

LeEtta Schmidt, lmschmidt@usf.edu and Drew Smith dsmith@usf.edu
Reviewed by USF General Counsel 08/11/2015
Overall, the amount and substantiality of material used in relation to the whole supports fair use or does not support fair use.

**EFFECT ON THE MARKET FOR ORIGINAL**

<table>
<thead>
<tr>
<th>Likely Supports Fair Use</th>
<th>Likely Does Not Support Fair Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>■ No significant effect on the market or potential market for the original</td>
<td>□ Replaces sale of copyrighted work</td>
</tr>
<tr>
<td>■ No similar product marketed by the copyright holder</td>
<td>□ Significantly impairs market or potential market for the work</td>
</tr>
<tr>
<td>■ You own a lawfully acquired copy of the material</td>
<td>□ Numerous copies or repeated, long-term use</td>
</tr>
<tr>
<td>□ The copyright holder is unidentifiable</td>
<td>■ Made accessible on Web or to public</td>
</tr>
<tr>
<td>□ Lack of licensing mechanism for the material</td>
<td>□ Affordable and reasonably available permissions or licensing</td>
</tr>
</tbody>
</table>

Overall, the effect on the market for the original supports fair use or does not support fair use.

**CONCLUSION**

The combined purpose and character of the use, nature of the copyrighted material, amount and substantiality of material used in relation to the whole and the effect on the market for the original supports fair use or does not support fair use.

*Note: Should your use of copyrighted material not support fair use, you may still be able to locate and request permissions from the copyright holder. For help on this, please feel free to contact your Copyright Librarian.*

This worksheet has been adapted from:

**Cornell University’s Checklist for Conducting A Fair use Analysis Before Using Copyrighted Materials:**
https://copyright.cornell.edu/policies/docs/Fair_Use_Checklist.pdf


**Smith, Kevin; Macklin, Lisa; A.; Gilliland, Anne. A Framework for Analyzing any Copyright Problem.** Retrieved from:
https://d395qusza40rc.cloudfront.net/cfel/Reading%20Docs/A%20Framework%20for%20Analyzing%20Arbitrary%20Copyright%20Problem.pdf

LeEtta Schmidt, lmschmidt@usf.edu and Drew Smith dsmith@usf.edu
Reviewed by USF General Counsel 08/11/2015