

UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Computational drug design applied to neglected disease

Permalink

<https://escholarship.org/uc/item/6kg5q2w8>

Author

Durrant, Jacob Devin

Publication Date

2010

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA, SAN DIEGO

Computational Drug Design Applied to Neglected Disease

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

in

Biomedical Sciences

by

Jacob Devin Durrant

Committee in charge:

Professor J. Andrew McCammon, Chair
Professor Philip Bourne
Professor Partho Ghosh
Professor Victor Nizet
Professor Sharon Reed

2010

Copyright

Jacob Devin Durrant, 2010

All rights reserved.

The Dissertation of Jacob Devin Durrant is approved, and it is acceptable in quality
and form for publication on microfilm and electronically:

Chair

University of California, San Diego

2010

DEDICATION

To my family for their dedicated support.

EPIGRAPH

...particles in nature are continually jiggling and bouncing, turning and twisting around one another. You will have to imagine this as a dynamic rather than a static picture...
The atoms are not standing still. They are jiggling and vibrating.

Richard P. Feynman, 1965 Nobel Laureate

TABLE OF CONTENTS

SIGNATURE PAGE	iii
DEDICATION	iv
EPIGRAPH.....	v
TABLE OF CONTENTS	vi
LIST OF ABBREVIATIONS	viii
LIST OF FIGURES	xi
LIST OF TABLES	xiii
ACKNOWLEDGEMENTS	xiv
VITA.....	xix
ABSTRACT OF THE DISSERTATION.....	xxi
CHAPTER 1: Computational Drug Design Applied to Neglected Disease.....	1
1.1 Introduction	1
1.2 Background and Significance of Selected Drug Targets.....	2
1.3 Established Methodologies: Molecular Dynamics Simulations and the Relaxed Complex Scheme.....	4
1.4 Developing New Drug-Discovery Methodologies.....	6
CHAPTER 2: Novel Inhibitors of <i>Trypanosoma Brucei</i> UDP-Glucose 4'-Epimerase Identified through Virtual Screening into a Flexible Protein	10
2.1 Introduction	10
2.2 Materials and Methods	13
2.3 Results/Discussion.....	20
2.4 Conclusion.....	36
2.5 Acknowledgements section	37
CHAPTER 3: Novel Naphthalene-Based Inhibitors of <i>Trypanosoma Brucei</i> RNA Editing Ligase I	39
3.1 Introduction	39
3.2 Materials and Methods	41
3.3 Results/Discussion.....	45
3.4 Conclusion / Future Directions.....	55
3.5 Acknowledgements	58
CHAPTER 4: Potential Drug-Like Inhibitors of Group 1 Influenza Neuraminidase Identified through Computer-Aided Drug Design	59
4.1 Introduction	59
4.2. Material and Methods.....	60
4.3 Results and Discussion	67
4.4 Conclusion	79

4.5 Acknowledgements	80
CHAPTER 5: Computational Identification of Uncharacterized Cruzain Binding Sites	82
5.1 Introduction	82
5.2 Methods	84
5.3 Results / Discussion.....	89
5.4 Acknowledgements	108
CHAPTER 6: Toward Understanding the Conformational Dynamics of RNA Ligation	109
6.1 Introduction	109
6.2 Materials and Methods	114
6.3 Results and Discussion.....	120
6.4 Conclusions	139
6.5 Acknowledgements	141
CHAPTER 7: A Multidimensional Strategy to Detect Polypharmacological Targets in the Absence of Structural and Sequence Homology	142
7.1 Introduction	142
7.2 Results	146
7.3 Discussion.....	150
7.4 Materials and Methods	160
7.5 Acknowledgements	164
CHAPTER 8: AutoGrow: A Novel Algorithm for Protein Inhibitor Design.....	165
8.1 Introduction	165
8.2 Methods and Materials	167
8.3 Results/Discussion.....	178
8.4 Acknowledgements	189
CHAPTER 9: Including Receptor Flexibility and Induced Fit Effects into the Design of MMP-2 Inhibitors.....	190
9.1 Introduction	190
9.2 Computational Details	196
9.3 Results and Discussions	200
9.4 Conclusion	211
9.5 Acknowledgments	211
References	213

LIST OF ABBREVIATIONS

Å	angstrom (10^{-10} meters)
ADME	absorption, distribution, metabolism, and excretion
ADP	adenosine diphosphate
ADT	AutoDockTools
AICAR	aminoimidazole carboxamide ribonucleotide
AIDS	acquired immune deficiency syndrome
AMP	adenosine monophosphate
AMPCPP	an unreactive analog of AMP
AMZ	aminoimidazole 4-carboxamide ribonucleotide
AppN	AMP attached to the nick 5'PO ₄ moiety
ATP	adenosine triphosphate
BSA	bovine serum albumin
CLogP	computed LogP
CoA	coenzyme A
DMSO	dimethyl sulfoxide
DNA	deoxyribonucleic acid
dsRNA	double-stranded ribonucleic acid
DTP	Developmental Therapeutics Program
<i>Efligase</i>	<i>Enterococcus faecalis</i> v583 NAD-dependent DNA ligase
FAS	fatty acid synthesis
FDA	Food and Drug Administration

FKBP	FK506 binding protein
gRNA	guide RNA
HAT	human African trypanosomiasis
HIV	human immunodeficiency virus
<i>HsETR1</i>	<i>Homo sapiens</i> mitochondrial 2-enoyl thioester reductase
<i>HsGalE</i>	<i>Homo sapiens</i> UDP-galactose 4-epimerase
<i>HsLigI</i>	<i>Homo sapiens</i> DNA ligase I
<i>HsLigIIIβ</i>	<i>H. sapiens</i> DNA ligase III β
<i>HsPDE9A2</i>	<i>Homo sapiens</i> phosphodiesterase 9A
IC50	half maximal inhibitory concentration
K_i	dissociation constant
LogP	partition coefficient
MD	molecular dynamics
MMP	matrix metalloproteinase
mRNA	messenger ribonucleic acid
<i>Mt</i> ligase	<i>Mycobacterium tuberculosis</i> DNA ligase
NAD	nicotinamide adenine dinucleotide
NADPH	nicotinamide adenine dinucleotide phosphate
NAMD	nanoscale molecular dynamics
NCBI	National Center for Biotechnology Information
NCI	National Cancer Institute
NMR	nuclear magnetic resonance

PCA	principal component analysis
PDB	protein data bank
PME	particle mesh ewald
RESP	restrained electrostatic potential
RMSD	root mean square deviation
RMSF	root mean square fluctuation
RNA	ribonucleic acid
SCOP	structural classification of proteins
<i>SeRmlB</i>	<i>Salmonella enterica</i> dTDP-D-glucose 4,6 dehydratase
SOIPPA	sequence order independent profile-profile alignment
<i>SpPce</i>	<i>Streptococcus pneumoniae</i> teichoic acid phosphorylcholine esterase
T4 Rnl1	T4 RNA ligase 1
T4 Rnl2	T4 RNA ligase 2
<i>TbREL1</i>	<i>Trypanosoma brucei</i> RNA editing ligase 1
<i>TbREL2</i>	<i>Trypanosoma brucei</i> RNA editing ligase 2
<i>TbGalE</i>	<i>Trypanosoma brucei</i> UDP-galactose 4' epimerase
UTP	uridine triphosphate
VMD	Visual Molecular Dynamics
XMP	xanthosine-5'-monophosphate
ZBG	zinc-binding group
ZINC	ZINC is not commercial

LIST OF FIGURES

Figure 1: The predicted binding interactions of clorobiocin bound to TbGalE	27
Figure 2: The predicted binding modes of compounds 1 and 2 to TbGalE	30
Figure 3: <i>TbREL1</i> inhibitors	42
Figure 4: The <i>TbREL1</i> active site	46
Figure 5: Binding modes of identified <i>TbREL1</i> inhibitors	50
Figure 6: Role receptor flexibility plays in <i>TbREL1</i> binding.....	55
Figure 7: The top four predicted neuraminidase inhibitors	76
Figure 8: Cruzain MD equilibration	90
Figure 9: The cruzain active site	91
Figure 10: The binding of compound 1 to the cruzain semi-open conformation	93
Figure 11: An uncharacterized binding pocket adjacent to the cruzain S2 subsite	95
Figure 12: The opening and closing of the two cruzain “gates”	97
Figure 13: Cruzain colored by residue conservation	102
Figure 14: Nicked dsRNA bound to TbREL1, poised to catalyze religation	113
Figure 15: The equilibrated TbREL1 structure	121
Figure 16: The TbREL1 ribofuranose pseudorotation angle.....	126
Figure 17: Equilibrated coordination of the TbREL1 active-site magnesium ion	128
Figure 18: RNA binding footprint.....	132
Figure 19: RNA flexibility	135
Figure 20: The AutoGrow strategy workflow	145
Figure 21: An illustration of the AutoGrow mutation and crossover operators.....	169
Figure 22: Default fragment libraries used by the AutoGrow mutation operator	171
Figure 23: The AutoGrow large- and small-fragment libraries.....	172
Figure 24: New fragment libraries to recreate ATP, oseltamivir, and AMZ	175
Figure 25: Cocrystallized and AutoGrow-generated ATP docked into TbREL1	181
Figure 26: Cocrystallized and AutoGrow-generated oseltamivir-like compounds	183
Figure 27: Cocrystallized and AutoGrow-generated AMZ.....	184
Figure 28: The MD trajectory of the MMP-2 apo protein.....	192
Figure 29: The workflow used to design novel MMP-2 inhibitors	201

Figure 30: LUDI-docked fragments added to a ZBG.....	210
Figure 31: Induced-fit contribution to the binding energy of MMP-2 inhibitors.....	204

LIST OF TABLES

Table 1: Fourteen low-micromolar <i>Tb</i> GalE inhibitors	28
Table 2: Affinities of Identified <i>Tb</i> REL1 inhibitors.....	49
Table 3: AutoGrow-generated compounds prior to modification	64
Table 4: Common fragments derived by decomposing FDA-approved drugs.....	70
Table 5: Four predicted potent neuraminidase inhibitors	75
Table 6: <i>Tb</i> REL1 active-site interactions	122
Table 7: RNA- <i>Tb</i> REL1 interactions.....	134
Table 8: Selected predicted secondary targets in humans.	148
Table 9: Selected predicted secondary targets in pathogens.	149
Table 10: Homology between <i>Hs</i> LigI and three other selected DNA ligases.....	153
Table 11: MMP-2 inhibitor candidates.....	193
Table 12: Active-site structural changes that occur upon ligand binding	206

ACKNOWLEDGEMENTS

First and foremost, I'd like to thank my PhD advisor, Dr. J. Andrew McCammon, and his computational biology group at the University of California, San Diego. Dr. McCammon has created a most remarkable lab environment that provided me with a unique and rewarding PhD experience. Principal investigators often err by either giving their students too much freedom or by micromanaging their every move. Dr. McCammon has found the perfect balance; he's very approachable and provides direction, assistance, and support when needed, while simultaneously allowing his students to explore their own ideas.

For all his academic accomplishments, the thing that initially impressed me about Dr. McCammon is his humanity. I met Dr. McCammon for the first time when I interviewed at UCSD prior to starting my MD/PhD degree. Andy sat me down and began to describe his research, focusing particularly on the role he played in the development of the first HIV protease inhibitor. A simple graph showed the number of AIDS deaths each year; following the discovery of the HIV protease inhibitor, the number of deaths dropped precipitously. He explained that the lives of tens of thousands, if not hundreds of thousands, had been and would be saved thanks to these novel therapeutics. While this statistic was impressive and certainly caught my attention, Andy's tone was also memorable. Despite so incredible an accomplishment, he spoke without a hint of arrogance, almost, if I read him right, with a reverent respect at having been able to participate in so important a project. In the world of academics, big egos are all too commonplace. Andy's approach to science is inspiring.

Numerous others in the lab have also been great mentors and colleagues. I first worked with Dr. César Augusto F. de Oliveira, a post-doc who helped me gain further experience with molecular dynamics simulations. We've had many discussions about protein flexibility and its role in ligand binding since that initial project, and I've learned a lot from him. Dr. Rommie Amaro, now on the faculty of the University of California at Irvine, has also been a wonderful mentor. She taught me about virtual screening, a technique critical to the research I've performed as a PhD student. Her patience and diplomacy are truly remarkable. Many conversations with others, including Elaine Thompson, Anthony Ivetaç, Susan Taylor, and William Sinko, have also been most educational.

I'd also like to thank Patricia Craft, the administrative assistant and nerve center of the McCammon lab who, as best I can tell, is able to solve any conceivable problem, no matter how difficult. I've also had a host of excellent collaborators, including Lei Xie, Michael Urbaniak, Michael Ferguson, Antti Haapalainen, Zhijun Chen, Anne Marie di Guilmi, Frank Wunder, Philip Bourne, and Robert Swift, among others. My doctoral committee has also provided excellent direction and counsel. Aside from Dr. McCammon, I'd also like to thank Drs. Victor Nizet, Philip Bourne, Partho Ghosh, and Sharon Reed, who likewise served on my committee.

Key undergraduate academic advisors who encouraged me to pursue a career in science also deserve mention. Dr. Daniel Fairbanks, the Dean of Undergraduate Education at Brigham Young University during my time there, has been a great inspiration to me. When I returned from Brazil, where I worked as a two-year

volunteer clergyman, Dr. Fairbanks took me under his wing. Under the influence of his mentorship, a vague desire to help the poor I'd met in South America was transformed into an organized plan to pursue an MD/PhD and to study neglected tropical diseases. The impact he's had on my life is incalculable.

Dr. David D. Busath also had an instrumental influence in my personal and professional development. When I spammed BYU's biology department looking for potential projects, Dr. Busath responded enthusiastically. I initially ignored his email, judging his computational project to be too difficult, but he persistently spammed me back, encouraging me to undertake the proposed project. I eventually joined his lab, where he introduced me to molecular and mathematical modeling and computational biology. How my life would be different had I ignored his persistent emails!

My family has also been a great source of support. So much of who I am I owe to my parents' many years of dedicated guidance. My wife Taciana has also been there for me during good times and bad. I love her dearly.

Authorship Acknowledgements:

Chapter 2, in full, is a near reprint of material that was recently accepted for publication in the Journal of Medicinal Chemistry. The dissertation author and Michael D. Urbaniak were the primary co-contributors to this work. Michael A. J. Ferguson and J. A. McCammon served as the advisers and principal investigators. The Journal of Medicinal Chemistry does not consider academic theses to constitute prior publication and so implicitly authorizes the inclusion of this chapter.

Chapter 3 is in large part a reprint of the material as it appears in a manuscript recently submitted to PLoS Neglected Tropical Diseases. The dissertation author and Laurence Hall were the primary co-contributors to this work. Achim Schnauffer and Rommie E. Amaro served as advisers and principal investigators. PLoS journals have an open-access policy that permits republication of PLoS articles in academic theses.

Chapter 4, in full, is a near reprint of the material as it appears in Computational Biology and Chemistry, volume 34, pages 97-105. The dissertation author was the primary co-contributor to this work. J. Andrew McCammon served as the head adviser and principal investigator. Elsevier holds the copyright for this manuscript and has authorized its publication in this dissertation.

Chapter 5, in full, is a near reprint of the material as it appears in PLoS Neglected Tropical Diseases, volume 4, issue 5, e676. The dissertation author and Henrik Keränen were the primary co-contributors to this work. Benjamin A. Wilson also made significant contributions, and J. Andrew McCammon served as head adviser and principal investigator. PLoS journals have an open-access policy that permits republication of PLoS articles in academic theses.

Chapter 6, in full, is a near reprint of the material as it appears in Biochemistry 2009, volume 3, pages 709-719. The dissertation author and Robert V. Swift were the primary co-contributors to this work. Rommie E. Amaro also made significant contributions, and J. Andrew McCammon served as head adviser and principal investigator. The American Chemical Society holds the copyright for this manuscript and has authorized its publication in this dissertation.

Chapter 7, in full, is a near reprint of the material as it appears in PLoS Computational Biology 2010, volume 6: e1000648. The dissertation author was the primary contributor to this work. Rommie E. Amaro, Lei Xie, Michael D. Urbaniak, Michael A. J. Ferguson, Antti Haapalainen, Zhijun Chen, Anne Marie Di Guilmi, Frank Wunder, and Philip E. Bourne also made significant contributions, and J. Andrew McCammon served as head adviser and principal investigator. PLoS journals have an open-access policy that permits republication of PLoS articles in academic theses.

Chapter 8, in full, is a near reprint of the material as it appears in Chemical Biology and Drug Design 2009, volume 73, pages 168-178. The dissertation author was the primary contributor to this work. Rommie E. Amaro also made significant contributions, and J. Andrew McCammon served as head adviser and principal investigator. Jacob Durrant holds the copyright for this manuscript and has authorized its publication in this dissertation.

Chapter 9, in full, is a near reprint of the material as it appears in Journal of Molecular Recognition 2009, volume 23, issue 2, pages 173-182. The dissertation author was the primary contributor to this work. César Augusto F. de Oliveira also made significant contributions, and J. Andrew McCammon served as head adviser and principal investigator. The Journal of Molecular Recognition holds the copyright for this manuscript and has authorized its publication in this dissertation.

VITA

1998 – 1999, 2001 – 2005	Bachelor of Arts, Portuguese Language and Literature Brigham Young University
1998 – 1999, 2001 – 2005	Bachelor of Science, Mathematics Brigham Young University
2004 – 2005	Master of Science, Physiology and Developmental Biology Brigham Young University Research Mentor: Prof. David Busath
2005 – 2007	Medical School University of California at San Diego
2007 – 2010	Graduate Student Researcher University of California at San Diego Biomedical Sciences Graduate Program Adviser: Prof. J. Andrew McCammon
2007 – 2010	Doctor of Philosophy, Biomedical Sciences University of California at San Diego Dissertation: “Computational Drug Design Applied to Neglected Disease”

PUBLICATIONS

1. *Computational Identification of Uncharacterized Cruzain Binding Sites*. Jacob D. Durrant; Henrik Keränen; Benjamin Wilson; J. Andrew McCammon. Recently accepted by PLoS Neglected Tropical Diseases.
2. *A Multidimensional Strategy to Detect Polypharmacological Targets in the Absence of Structural and Sequence Homology*. Jacob D. Durrant; Rommie E. Amaro; Lei Xie; Michael D. Urbaniak; Michael A. J. Ferguson; Antti Haapalainen; Zhijun Chen; Anne Marie Di Guilmi; Frank Wunder; Philip E. Bourne; J. Andrew McCammon. 2010. PLoS Comp. Bio., 6(1):e1000648.
3. *Including receptor flexibility and induced fit effects into the design of MMP-2 inhibitors*. J. Durrant; C. A. F. Oliveira; J. A. McCammon. 2009. J. Mol. Recognition, Early View
4. *Toward Understanding the Conformational Dynamics of RNA Ligation*. Swift, R. V.; Durrant, J.; Amaro, R. E.; McCammon, J. A. 2009. Biochemistry, 48(4):709–19.
5. *AutoGrow: A Novel Algorithm for Protein Inhibitor Design*. Durrant, J. D.; Amaro, R. E.; McCammon, J. A. 2009. Chemical Biology & Drug Design. 73(2):168-78.
6. *Tryptophan contributions to the empirical free-energy profile in gramicidin A/M heterodimer channels*. Durrant, J. D.; Caywood, D.; Busath, D. D. 2006. Biophys J. 1;91(9):3230-41.
7. *An algorithm for analyzing linkages affected by heterozygous translocations: QuadMap*. Durrant, J. D.; Gardunia, B. W., Livingstone, K. D., Stevens, M. R.,

- Jellen, E. N. 2006. *Journal of Heredity* 97:62-66.
8. *Modulation of Lipid Bilayer Interfacial Dipole Potential by Phloretin, RH421, and 6-Ketocholestanol as Probed by Gramicidin Channel Conductance.* Duffin, R. L.; Garrett, M. P.; Flake, K. B.; Durrant, J. D.; Busath, D. D.; *Langmuir*; 2003; 19(8); 3561-3561.
 9. *Segregation Patterns in Oat Populations Derived From Parents Heterozygous for Translocations 7C-17 and 6C-21.* Eric N. Jellen, Brian W. Gardunia, Jacob Durrant, F. Douglas Raymond, J. Paul Murphy, David P. Livingston, and Albert G. Santos. 2002. American Oat Workers Conference Proceedings.

ABSTRACT OF THE DISSERTATION

Computational Drug Design Applied to Neglected Disease

by

Jacob Devin Durrant

Doctor of Philosophy in Biomedical Sciences

University of California, San Diego, 2010

Professor J. Andrew McCammon, Chair

Herein is described the identification of novel inhibitors of key pathogenic proteins, the use of established computational techniques to further drug discovery, and the development of novel computational methods, with the ultimate goal of identifying small-molecule compounds that, with further development, may serve as future treatments for neglected diseases. Two of the pathogens targeted, *Trypanosoma brucei* and *Trypanosoma cruzi*, are neglected because they are tropical, and drugs

marketed solely to the developing world are rarely profitable. A third disease, H1N1 pandemic influenza, is somewhat neglected because it is only now emergent.

For *Trypanosoma brucei*, two drug-discovery projects are described, focusing on two crucial enzymes: UDP-galactose 4'-epimerase and RNA editing ligase 1. For *Trypanosoma cruzi*, the dynamics of cruzain, a crucial cysteine protease, are studied and characterized. Finally, for influenza, predicted inhibitors of neuraminidase (N1) are presented.

Aside from describing identified enzyme inhibitors, the current work also describes several new methods that may be generally applicable to drug discovery. These new methods include a multidimensional strategy for the identification of secondary targets of known small-molecule inhibitors in the absence of global structural and sequence homology with the primary target protein, a novel computer-aided drug design algorithm (AutoGrow) that combines the strengths of both fragment-based growing and docking algorithms, and an extension to the relaxed complex scheme that accounts for both population-shift and induced-fit contributions to ligand binding.

CHAPTER 1: Computational Drug Design Applied to Neglected Disease

Jacob D. Durrant

1.1 Introduction

Neglected diseases like malaria, leishmaniasis, lymphatic filariasis, Chagas' disease, and schistosomiasis continue to cause significant morbidity and mortality in developing nations, despite recent advances in the development of novel drug-discovery methods. Because developing drugs against diseases that primarily afflict the poor is rarely profitable, international pharmaceutical companies have largely abandoned any efforts to develop new therapies to fight neglected diseases.¹ This thesis describes work done to in part address this deficiency. The document is divided into three parts:

1. Chapters two to four describe the identification of novel inhibitors of key pathogenic proteins, including UDP-galactose 4'-epimerase and RNA editing ligase 1 from *Trypanosoma brucei*, as well as influenza N1 neuraminidase.
2. Chapters five and six describe the use of computational techniques, including molecular dynamics simulations and the identification of phylogenetically conserved residues, to further drug discovery.
3. Chapters seven to nine describe the development of novel methods to further computational drug discovery.

1.2 Background and Significance of Selected Drug Targets

Much of the research described centers on four critical proteins from three pathogenic species: *T. brucei* UDP-galactose 4' epimerase, *T. brucei* RNA editing ligase 1, influenza (N1) Neuraminidase, and *T. cruzi* cruzain.

T. brucei UDP-galactose 4' epimerase (*TbGalE*): *Trypanosoma brucei* is a fatal parasite for which drug development has been largely neglected.² Due to the weaknesses of available therapies, new therapies are critical to future treatment. Currently, there are 300,000 new cases of *T. brucei* infection per year.³ *TbGalE*, an enzyme of the Leloir pathway of galactose metabolism described in chapter two, is an excellent *T. brucei* drug target. In *T. brucei*, galactose is a component of the variant surface glycoproteins used for immune-system evasion,⁴ of the transferrin receptor used for iron acquisition,⁵ and of the parasite endocytic pathway.⁵ Curiously, while *T. brucei* has a hexose transporter capable of glucose uptake, it is unable to acquire galactose from the host^{6, 7}; intracellular galactose must be synthesized from glucose via *TbGalE*. Consequently, inhibition of this key step of trypanosomal galactose synthesis renders the parasite inviable.

T. brucei RNA editing ligase 1 (*TbREL1*): Another potential *T. brucei* drug target is RNA editing ligase 1 (*TbREL1*), described in detail in chapters three, six, and seven. In *T. brucei*, some mitochondrial RNA transcripts must be modified before they are translated. Following RNA editing, *TbREL1* religates nicked RNA, preparing it for the ribosome. As *TbREL1* is essential for *T. brucei* survival, and because there are no close *TbREL1* human homologues, *TbREL1* is an ideal drug target.

Influenza (N1) Neuraminidase: Due to the recent H1N1 pandemic, influenza has received a lot of attention in recent months. Aside from H1N1, human outbreaks of the H5N1 strain have already occurred in Hong Kong, Indonesia, Cambodia, Thailand, and Vietnam. Of even greater concern, some strains are resistant to antivirals like amantadine and oseltamivir.^{8,9} The need for novel antivirals is critical. Neuraminidase, described in chapter four, is an influenza surface protein essential for viral release. After viral particle formation, viruses remain bound to the cell surface *via* membrane-bound sialic-acid residues. Neuraminidase cleaves these sialic-acid residues, releasing the virus from the cell surface and allowing viral propagation. Because neuraminidase mutates frequently, every year brings news strains, necessitating new vaccines and, potentially, new neuraminidase inhibitors.¹⁰

T. cruzi cruzain: American trypanosomiasis, also known as Chagas disease, is endemic to Central and South America, where 50,000 die annually.¹¹ The disease is caused by the unicellular parasite *Trypanosoma cruzi* (*T. cruzi*), an organism transmitted by insects of the *Reduviidae* family. The major cysteine protease of *T. cruzi*, called cruzain or, alternatively, cruzipain, described in chapter five, is one attractive drug target.¹² A member of the peptidase C1 protein family, cruzain is present and essential in all stages of *T. cruzi* development.^{13, 14} Over-expression of cruzain enhances the transformation of the parasite into the infective form,¹⁵ and reduced protease activity prevents infection in wild-type mice.¹⁶ Additionally, cysteine protease inhibitors block both the replication and the differentiation of the parasite *in*

vitro and *in vivo*.^{12, 17-24} Cruzain inhibitors can cure infection in cell, mouse, and dog models.^{20, 25}

1.3 Established Methodologies: Molecular Dynamics Simulations and the Relaxed Complex Scheme

Established methodologies like molecular dynamics simulations and virtual screening drove much of the initial drug discovery described in this dissertation. The following paragraphs describe both of these methodologies in greater detail.

Molecular Dynamics Simulations

Classical molecular dynamics (MD) simulations, described in chapters two, three, four, five, six, and nine, seek to predict nanometer-scale molecular motions on the nanosecond to microsecond time scale. The study of molecular motions can provide important insights into protein-drug interactions.

In order to simplify the calculations, the forces acting on atoms are divided into two categories: bonded and nonbonded. Bonded forces are those that arise when atoms are chemically bonded to each other. In classical MD simulations, these forces are approximated by assuming that the atoms behave as balls connected by a complex system of springs. For example, in order to assure that two bonded atoms are appropriately separated, a spring with the appropriate stiffness is imagined to connect them, where the ideal spring length is equal to the ideal bond length. The angles between three connected atoms, as well as the dihedral angles between four connected

atoms, are likewise maintained by assuming that atomic movement is dictated by miniscule, imagined spring systems. While the spring approximation is not exact, it is far less computationally demanding than more precise quantum-mechanics based simulations, which at present can only be used to study small systems.

MD simulations also account for two kinds of nonbonded forces, van der Waals and electrostatic, which act on pairs of atoms that are not chemically bonded to each other. Van der Waals forces are approximated by an energy function, usually the Leonard-Jones potential, and electrostatic forces are calculated according to Coulomb's law. As both of these forces approach zero as distance increases, they are typically calculated only for nonbonded atoms that are physically close to each other in order to reduce the number of calculations needed.

Once the specific bonded and nonbonded forces are calculated, atom positions are updated by integrating Newton's equations of motion over a small time step, typically at most a few femtoseconds. Once the atoms of the system have been moved, the forces are again evaluated, and the algorithm continues iteratively until the simulation is stopped.

Virtual Screening

Aside from MD simulations, virtual screening is also a critical computational component of the drug-discovery projects described in this dissertation. Traditional biochemical screens can be very expensive; consequently, drug companies and academicians are turning increasingly to preliminary virtual screens to aid in the

identification of pharmacologically active compounds. Virtual screens utilize docking software together with a library of small-molecule compounds; the software attempts to position library compounds within a protein active site and to predict binding energy. Those compounds with high predicted affinities are then tested experimentally.

A number of free docking programs are available, including AutoDock 4.0²⁶ and AutoDock Vina.²⁷ Unfortunately, these docking algorithms do not typically take into account full protein receptor flexibility, leading to many false-positive hits. The relaxed complex scheme,²⁸ an advanced docking methodology described in chapters two and three, in part compensates for this weakness. The scheme begins with a MD simulation of the target protein. Following the simulation, each candidate inhibitor is docked into multiple representative protein configurations extracted from the MD trajectory, yielding a spectrum of binding energies for each compound. An ensemble-based score is then calculated for each compound that accounts for the many protein conformations sampled during the MD simulation. The relaxed complex scheme has already been used to successfully identify inhibitors of FKBP,²⁹ HIV integrase,³⁰ and *T. brucei* RNA editing ligase 1.³¹

1.4 Developing New Drug-Discovery Methodologies

Aside from using well established computational methodologies to identify novel protein inhibitors, my PhD studies also included the development of several

novel methodologies. Three of these novel methodologies, described in detail in chapters seven, eight, and nine, are introduced in the following paragraphs.

Identifying Secondary Targets: Side-Effect Prediction and Drug Retooling

Chapter seven describes the development and application of a novel algorithm to identify secondary targets of known inhibitors. While the pharmaceutical industry makes great efforts to ensure that drug binding is specific to a particular target protein, often times drugs bind other, unidentified proteins, called secondary targets, as well. By identifying secondary human targets, researchers could predict potential side effects and even modify lead drug candidates so as to reduce promiscuous binding, thereby avoiding the high costs associated with waiting until clinical trials to discover potential intolerable side effects. Indeed, unexpected side effects cause nearly one third of all drugs to fail.³²

Additionally, by identifying secondary targets in human pathogens, researchers could predict secondary therapeutic uses of known inhibitors. The retooling of drugs previously approved for other conditions may provide critical future therapies for neglected diseases. As designing novel drugs specifically to treat these diseases is rarely profitable, large pharmaceutical companies have by and large turned to more lucrative projects. To demonstrate the utility of this novel methodology, we identified several secondary targets of a known *Tb*REL1 inhibitor.

AutoGrow

AutoGrow, described in chapters four and eight, is a novel computer program that uses a fragment-based evolutionary algorithm to “evolve” drug candidates. For each generation, all ligand files are docked into the target protein and scored according to their predicted binding affinities. Those ligands with the best predicted binding affinities are selected for inclusion in the next generation. As generation after generation is created, each based on the best binding individuals of the previous generation, potent predicted inhibitors eventually evolve.

Further Studies of Relaxed Complex Docking

As described above, the relaxed complex scheme is a well established inhibitor-identification protocol. However, additional studies are needed to refine the method and increase its power. Chapter nine describes such a refinement.

Traditionally, two models have been used to explain ligand-protein interactions. The first, called the induced-fit model, suggests that in the absence of the ligand the protein does not take on a configuration amenable to ligand binding. Rather, the ligand itself induces changes in the receptor conformation, causing the active site to accommodate through rearrangement.

The second model, called the population-shift model, suggests that a protein is constantly transitioning between multiple conformational states. Only a certain subpopulation of the states is amenable to ligand binding. When a ligand does bind, it stabilizes the amenable conformations, and the population of conformations

subsequently shifts. In reality, ligand binding probably occurs *via* a combination of these two mechanisms.

The relaxed complex scheme accounts for population-shift effects by docking candidate inhibitors into diverse static structures extracted from a dynamic simulation. However, it does not necessarily account for induced-fit effects, possible changes in the active-site conformation that occur only in the presence of the ligand.

In chapter nine, an extension to the relaxed complex methodology is described. As before, candidate compounds are docked into structures extracted from an MD simulation. However, additional MD simulations are then performed of the top-scoring, docked relaxed-complex compounds to account for changes in the protein conformation that only occur in the presence of the ligand. Importantly, accounting for induced-fit effects significantly improved the ranking of several known inhibitors (positive controls).

CHAPTER 2: Novel Inhibitors of *Trypanosoma Brucei* UDP-Glucose 4'-Epimerase Identified through Virtual Screening into a Flexible Protein

Jacob D. Durrant, Michael D. Urbaniak, Michael A. J. Ferguson, and J. Andrew McCammon

2.1 Introduction

Human African trypanosomiasis (HAT), a condition caused by Tsetse-fly mediated transmission of the unicellular parasite *Trypanosoma brucei*,³³ affects between 50,000 and 70,000 people living in sub-Saharan Africa, with an additional 50 million people at risk of infection.³⁴ The initial stage of the disease, called the hemolymphatic stage, is characterized by mild symptoms including fever, arthralgia, headache, and pruritus. Once the parasite crosses the blood-brain barrier, the neurological symptoms of the chronic stage of the disease are manifest, including severe headaches, nocturnal insomnia, mental and psychological disturbances, coma, and death if untreated.³⁴

Pentamidin and suramin, effective against the subspecies *T. b. gambiense* and *T. b. rhodesiense*, respectively, are typically administered during the early hemolymphatic stage.³⁵⁻³⁷ The approved treatments for advanced HAT, however, melarsoprol and eflornithine, are entirely inadequate. Melarsoprol causes encephalopathy in one out of every twenty recipients and is subject to growing drug resistance.^{34, 38, 39} Eflornithine is ineffective against the *T. b. rhodesiense* subspecies⁴⁰ and can cause fever, seizures, and infections, though combination therapy with nifurtimox ameliorates some of these side effects.⁴¹ To date, no vaccine is available

because *T. brucei* undergoes antigenic variation by altering its surface glycoprotein coat, thereby evading the immune system.⁴²

As drugs that target the infections of the developing world are rarely profitable, pharmaceutical companies have largely neglected the development of novel HAT therapeutics. Of the drugs listed above, for example, only one, eflornithine, has been developed since the late 1940s.⁴³ The trypanocidal effect of eflornithine was discovered only by accident, and it is only available today because the compound has also been commercialized as a cosmetic cream for the treatment of hirsutism.

This neglect provides researchers in academia with a unique opportunity to step in and address a largely unmet need. While a few academic institutions perform high-throughput screens to identify novel inhibitors of pathogenic enzymes, these large-scale projects are often cost prohibitive outside of industry. Fortunately, recent advances in computer-aided drug design have provided academic researchers with powerful tools that in part compensate for insufficient funding.

Computer docking constitutes one such advance.⁴⁴ Given a known protein structure from x-ray crystallography, homology modeling, or NMR, docking programs attempt to predict the binding mode and affinity of candidate inhibitors. For each candidate, a docking score is reported that, ideally, is proportional to the actual binding affinity. Researchers often dock thousands of compounds into protein structures of interest, effectively replicating virtually the kind of high-throughput screens so common in industry. While false positives and negatives are common,

computer docking can at least serve to enrich a pool of candidate compounds for likely inhibitors.

Traditional computer-docking methodologies often fail to identify true-positive inhibitors, however, because current algorithms do not typically account for full protein flexibility. When a ligand approaches its receptor *in vitro* or *in vivo*, it encounters not a static protein structure but rather an ensemble of structures as the protein “breathes” in solution. Ligand binding can potentially occur to any one of the many protein conformations sampled. The static protein structures typically used for computer docking studies do not capture the highly dynamic reality of ligand binding.

Recently, a new docking protocol called the relaxed complex scheme²⁸ has been developed that takes into account full protein flexibility. Rather than docking candidate compounds into a static protein receptor, compounds are docked into an ensemble of protein conformations typically extracted from an MD simulation. The compounds are then ranked by an ensemble-based score that accounts for active-site dynamics. The relaxed complex scheme has already been used to successfully identify inhibitors of FKBP,²⁹ HIV integrase,³⁰ and *T. brucei* RNA editing ligase 1.³¹

UDP-galactose 4' epimerase (*TbGalE*), a short-chain dehydrogenase / reductase enzyme of the Leloir pathway of galactose metabolism,^{45, 46} is one promising *T. brucei* drug target that has not yet been subjected to a relaxed complex screen. *TbGalE* is essential for parasite survival; null mutants of both the bloodstream and procyclic form of the parasite cannot survive.^{47, 48} In *T. brucei*, galactose is a component of the variant surface glycoproteins used for immune-system evasion,⁴ of

the transferrin receptor used for iron acquisition,⁵ and of the parasite endocytic pathway.⁵ Curiously, while *T. brucei* has a hexose transporter capable of glucose uptake, it is unable to acquire galactose from the host.^{6,7} Intracellular galactose must be synthesized from glucose *via* *TbGalE*.

H. sapiens UDP-galactose 4' epimerase (*HsGalE*) has 33% sequence identity with *TbGalE*.⁴⁵ Some drug specificity may be achieved by exploiting a non-conserved substitution from G237 in *HsGalE* to C266 in *TbGalE*.⁴⁵ Even if specificity for the *T. brucei* enzyme cannot be achieved, however, *TbGalE* may yet serve as a potential drug target. As human cells can import both glucose and galactose, inhibition of intracellular galactose synthesis is not lethal. Patients with a natural mutation in *HsGalE* develop a rare form of galactosaemia that can be partially controlled by dietary regulation.⁴⁵ Dietary regulation during anti-trypanosomal therapy may therefore be sufficient to prevent the potential side effects of a *TbGalE* inhibitor if drug specificity cannot be achieved.

Motivated by the urgent need for novel HAT therapeutics and the great potential of *TbGalE* as a drug target, we here use the relaxed complex scheme to identify 14 low-micromolar *TbGalE* inhibitors.

2.2 Materials and Methods

System Preparation

Chains A and B of a 2.0 Å resolution *TbGalE* crystal structure (PDB ID: 1GY8⁴⁵) were processed for use in a molecular dynamics simulation. All crystallographic

waters within 4 Å of the protein were retained, and missing loops were modeled using Discovery Studio 2.1 (Accelrys). Hydrogen atoms were added to the protein residues using the PDB2PQR server⁴⁹; histidine protonation states were manually verified. Additional hydrogen atoms were added to the ligand and co-factor using Discovery Studio 2.1 (Accelrys).

Force-field parameters for NAD⁺ were obtained from Richard B. Ryce.^{50, 51} Atom-type parameters for uridine diphosphate (UDP) were derived from the ff99 force field, based on the uridine monophosphate (RU) residue.⁵² The linking ribose oxygen atom of RU was replaced with a hydroxyl group, and the atom types of the UDP beta phosphate were chosen to be the same as those of the RU alpha phosphate. *Ab initio* partial charges were calculated for the UDP atoms using the GAUSSIAN03 computer program. The Hartree-Forck (HF) method, in conjunction with the 6-31+G* basis set, was used for geometry optimization. The optimized structure was used for subsequent single-point charge calculations performed using multi-conformational RESP fitting.⁵³ The remaining protein residues were parameterized using the ff99 force field.⁵²

The LEAP module of AMBER9 was subsequently used to further process the system.⁵⁴ The protein was submerged in a TIP3P water box⁵⁵ that extended 10 Å beyond the protein in the x, y, and z directions. Eighteen Na⁺ ions were added to bring the system to electrical neutrality. An additional 13 Na⁺ and Cl⁻ ions were added to simulate a 20 mM solution.

Molecular Dynamics (MD) Simulations

The system was relaxed *via* a four-phase minimization protocol. In the first phase, 5,000 steps of minimization were applied to the hydrogen atoms alone. In the second phase, 5,000 steps of minimization were applied to the hydrogen atoms, the water molecules, and all ions. Ten thousand steps of minimization were then applied to the hydrogen atoms, the water molecules, all ions, and the atoms of the protein backbone. Finally, all atoms were minimized for an additional 25,000 steps.

A four-phase equilibration protocol was next used. Starting from the minimized system, four sequential, 250-ps MD simulations at 310 K were executed with harmonic restraining forces of 4.0, 3.0, 2.0, and 1.0 kcal mol⁻¹ Å⁻² applied to the atoms of the protein backbone for the first, second, third, and fourth simulation, respectively.

This equilibrated system was used as the starting point for five distinct 20-ns, unrestrained, “productive” MD simulations. Because *TbGalE* is a homodimer, 200 ns of monomer dynamics were effectively simulated. A 2-fs time step was used. The temperature and pressure of the system were maintained by Langevin dynamics at 310 K and the hybrid Nose Hoover-Langevin piston method at 1 atm, respectively,⁵⁶ with period and decay times of 100 and 50 fs, respectively. Long-range electrostatics were treated using the particle mesh Ewald algorithm.⁵⁷ Bonded interactions were calculated at every time step. Nonbonded interactions and full electrostatics were calculated every two time steps. Bonds with hydrogen atoms were constrained using

the SHAKEH algorithm.⁵⁸ NAMD 2.6 was used to perform all minimizations, equilibrations, and productive simulations.⁵⁹

Conformational Clustering

To identify representative protein conformations sampled during the MD simulation, each of the ten effective trajectories (five simulations times two monomers per simulation) were concatenated, and the composite trajectory was aligned by minimizing the root-mean-square difference (RMSD) between the active-site atoms of each frame and the corresponding atoms of the first frame. RMSD clustering using the *gromos* algorithm as implemented in the GROMOS++ analysis software^{60, 61} was used to identify representative structures. A RMSD cutoff of 1.4 Å produced 24 clusters. The centroid member of each cluster was selected, and the set of all 24 centroids, representative of all the protein configurations sampled during the MD simulation, is said to constitute an *ensemble*.

Docking into the Crystal Structure

AutoDock Vina (Vina)²⁷ was used for all computer dockings, as that program was able to adequately recapture the crystallographic poses of the ligand and co-factor. Four virtual screens of the NCI Diversity Set II against a *Tb*GaIE crystal structure (PDB ID: 1GY8) were performed. For three of the four virtual screens, both the UDP ligand and the NAD co-factor were removed from the receptor, and the compounds of the NCI Diversity Set II were docked into the UDP binding site, into the NAD binding

site, and into both contiguous sites simultaneously, respectively. For the fourth virtual screen, the NAD co-factor was maintained, and the compounds of the NCI Diversity Set II were docked into the UDP binding site.

Relaxed Complex Scheme

In order to account for full protein receptor flexibility, the four screens performed against the crystal structure were repeated. Instead of docking into a static crystal structure, however, each compound of the NCI Diversity Set II was docked into the 24 representative receptor configurations of the ensemble. The compounds were ranked according to two ensemble-based criteria. The first was an ensemble-average binding energy, given by

$$\bar{E} = \frac{\sum_{i=1}^{23} w_i E_i}{\sum_{i=1}^{23} w_i} \quad (1)$$

where $\bar{E}$ is the weighted ensemble-average score, w_i is the size of cluster i , and E_i is the docking score of the ligand docked into the centroid of cluster i .

The second was the ensemble-best binding energy. For each of the compounds of the library, 24 predicted binding energies were calculated, one predicted energy for each of the ensemble protein conformations. The best binding energy of these 24 was identified, and the compounds were ranked by this ensemble-best score.

AutoDock 4.0 Rescoring

As Vina has never been used in a relaxed complex screen, each docked pose was rescored with the AutoDock 4.0 (AutoDock) scoring function,²⁶ without redocking. Ensemble-average and ensemble-best scores were recalculated based on the AutoDock 4.0 score as well, and compounds were reranked accordingly.

Similarity Searches

The entire NCI database was searched for compounds similar to those identified in the primary relaxed complex screen. Two methods were used to judge compound similarity. First, 19 compounds were identified that contained a 2'-(phenylcarbamoyl)-[1,1'-biphenyl]-2-carboxylic acid substructure. Second, 57 similar compounds were identified using a Tanimoto index with a 90% cutoff. Finally, four additional compounds similar to clorobiocin were identified using a Tanimoto index with an 80% cutoff.

LogP Calculation

To predict the LogP of selected compounds, the structures were first processed with LigPrep (Schrödinger) to generate various ionization ($\text{pH } 7.0 \pm 2.0$) and stereoisomeric states. The LogP of each of these structures was calculated using QikProp, also part of the Schrödinger suite. For those compounds with multiple ionization and stereoisomeric states, the associated LogP values were averaged to give one value per compound.

TbGalE Inhibition Assay

All compounds were obtained from the NCI/DTP Open Chemical Repository (<http://dtp.cancer.gov>). Compound identity was independently confirmed by high-resolution mass spectrometry, and LC-MS was used to confirm the purity >99%.

Recombinant *TbGalE* containing an N-terminal hexahistidine tag was expressed in *E. coli* and purified as described previously.⁴⁷ The inhibition of *TbGalE* was measured using high pH anion exchange chromatography (HPAEC) to follow the conversion of UDP-Gal to UDP-Glc by *TbGalE*.⁶² The reaction mixture (1 mM Tris pH 7.6, 100 μ M UDP-Gal, 100 μ M b-NAD⁺, 5 mg/mL *TbGalE*, 1% dimethyl sulphoxide) was incubated at 37 °C for 30 min with or without inhibitor (0.03 – 100 μ M), quenched with 15 μ L of 10 mM NaOH, and then subjected to HPAEC chromatography on a CarboPac PA-1 column (Dionex) in 1 mM NaOH with a NH₄OAc gradient.⁶² The eluant was monitored at 260nm, and peaks assigned by comparison to commercial standards. The IC₅₀ value was calculated using a four-parameter fit of the potency curves derived from three independent experiments.

To exclude the possibility of aggregation effects, the inhibition of the compounds was measured at 30 μ M ($\sim 5 \times \text{IC}_{50}$) in the presence and absence of a detergent (0.06% *n*-octylglucopyranoside), and no difference was observed.

Cytotoxicity Assay

Cytotoxicity assays were preformed against cultured *T. brucei* and human MRC5 cells as described previously,⁶² but further miniaturized to 384 well plates. Briefly, 50 μ l of cells (1×10^4 cells/ml) were aliquoted into 384-well plate and incubated with or without inhibitor (100 μ M to 30 nM) for a further three days before the number of viable cells was counted using the Alamar blue assay.⁶³

2.3 Results/Discussion

Human African trypanosomiasis threatens fifty million people in sub-Saharan Africa.³⁴ As current therapies are either too dangerous or too limited, novel drugs are urgently needed. UDP-galactose 4' epimerase (*TbGalE*), a protein critical to *T. brucei* survival, is one potential drug target. We here use computer-aided drug design to identify fourteen low-micromolar inhibitors of *TbGalE*.

Protein Dynamics

X-ray crystallography, which produces only static protein structures, cannot always capture the important role that protein dynamics plays in ligand binding.^{28, 64} Before attempting to identify potential *TbGalE* inhibitors, we first studied the molecular motions of *TbGalE*. Computer-aided drug design has traditionally neglected full protein flexibility. This neglect is surprising; many ligands bind to and stabilize rare active-site configurations. Upon binding, the population of active-site configurations sampled by the protein as it “breathes” in solution may shift to better

accommodate the ligand. Additionally, under the influence of a bound ligand, the protein may assume novel active-site configurations not sampled by the *apo* protein.⁶⁵

In order to better understand *TbGalE* dynamics, we performed five 20-ns molecular dynamics (MD) simulations. Five short simulations were chosen, as opposed to one long simulation, in order to increase the diversity of protein conformations sampled and also to ensure that the conformations sampled were geometrically similar to that of the known crystal structure. As *TbGalE* is a homodimer, each simulation provided two monomer trajectories, further increasing the diversity of conformations sampled. Clustering was subsequently used to identify 24 protein structures with active sites representative of the many active-site configurations sampled during the MD simulation. These 24 representative proteins are said to constitute an *ensemble*.

The Relaxed Complex Scheme

The protein conformations of the ensemble were easily incorporated into a virtual screening protocol called the relaxed complex scheme.²⁸ In a traditional virtual screen, the compounds of a small-molecule library are docked into a single static structure typically obtained from x-ray crystallography, homology modeling, or NMR. The compounds are then ranked by the corresponding docking score. In the relaxed complex scheme, each of the compounds of the library is docked into multiple protein receptor configurations typically extracted from a MD simulation. For each small molecule, an ensemble-based score is then calculated that accounts for the multiple

corresponding docking scores. Compounds are then reranked by this ensemble-based score, rather than by the single score associated with the crystal structure alone.

Several ensemble-based scoring schemes have been envisioned. First, the compounds can be ranked by the ensemble-average docking score. Scoring by the ensemble average seems reasonable, as such an average accounts for binding to each of the ensemble protein conformations and may minimize errors associated with docking to one structure alone. Alternatively, one could argue that the ensemble-best docking score more accurately captures the mechanics of natural ligand binding. All other things being equal, it is likely the protein conformation that best binds the ligand that would be most stabilized upon binding, and so it is that conformation that may be most relevant.

Another uncertainty centers on computer resource availability. As the relaxed complex scheme requires far more individual dockings than traditional virtual-screening protocols, the computational cost of docking all of the compounds of even a small library (~2000 compounds) into all the protein configurations of a moderately sized ensemble (~25 protein configurations) is prohibitive if a resource-intensive docking program like AutoDock 4.0 (AutoDock)²⁶ is used. Instead, a two-tiered protocol is typically employed. A traditional screen against the crystal structure initially filters out poor binders; the top hits from this initial screen are selected for further docking and scoring using the relaxed complex scheme.

However, compound ranking is often drastically altered when screened compounds are reranked by an ensemble-based score. While success has been reported

using the two-tiered method in the past,³¹ one weakness of that method is that some of the compounds omitted from the relaxed complex screen due to poor crystal-structure binding might in fact rank well as judged by an ensemble-based score. Consequently, we expect that the false negative rate would decrease if the initial crystal-structure screen used to filter out potential poor binders could be skipped.

The Primary Relaxed Complex Screen

The relaxed complex screen described in the current work was designed to address both these uncertainties. The compounds of the NCI Diversity Set II were ranked by both an ensemble-average and an ensemble-best scoring scheme so that the performance of both scoring criteria could be compared directly. Additionally, in hopes of reducing the number of false negatives, we eliminated the need for an initial crystal-structure screen by using AutoDock Vina (Vina),²⁷ a docking program that is two orders of magnitude faster than AutoDock. Vina performs well relative to AutoDock; while AutoDock is slightly better at predicting the energy of binding (standard error of 2.2 kcal mol⁻¹ versus 2.8 kcal mol⁻¹), Vina more accurately reproduces co-crystallized ligand poses.^{26, 27} To our knowledge, Vina has never been used in a relaxed complex screen.

Vina requires that the user define a docking grid (box) encompassing the active site. *Tb*GaIE binds both UDP-glucose and a NAD⁺ cofactor; potential inhibitors could occupy the binding pocket of either. To account for these two pockets, we ran four relaxed complex screens. In the first, the docking grid covered the UDP-glucose

binding pocket, and the NAD^+ cofactor was left intact. The NAD^+ cofactor was then removed for the second, third, and fourth screens, which had grids centered on the UDP-glucose, the NAD^+ , and the combined contiguous UDP-glucose / NAD^+ pockets, respectively.

Thus, the compounds of the NCI Diversity Set II were ranked in eight different ways; one of four docking grids was selected, and, for each grid, the compounds were ranked by either the ensemble-average or ensemble-best score. The top ten compounds from each of these virtual screens were combined into a single list of lead compounds; as there was significant redundancy, this single list contained 26 small molecules. When two positive controls (NAD^+ and UDP) were included in the screen, one or the other ranked among the top 30 inhibitors in all eight instances, attesting to the validity of the virtual-screening methodology used.

Rescoring the Primary Screen with AutoDock 4.0

As AutoDock is slightly better than Vina at predicting the energy of binding,^{26,}
²⁷ the docking poses of the eight relaxed complex screens described above were rescored using the AutoDock scoring function, without redocking. In six of the eight rescored screens, neither of the two positive controls ranked in the top 30; these six screens were discarded. The top ten compounds from each of the remaining two rescored screens (both, by chance, scored using the ensemble-best criteria) were merged into a single list. This list contained five compounds that were not among the

26 compounds identified using Vina alone. Thirty-one compounds were subsequently tested experimentally.

Experimental Validation Confirms Multiple Hits from the Primary Screen

Of the 31 compounds of the primary screen, eleven showed >50% average inhibition at 100 μ M. Interestingly, at this same concentration, seven compounds showed greater than two fold stimulation, suggesting allosteric cooperativity between the two monomers of the *TbGalE* homodimer, in harmony with previous studies that demonstrated GalE allostery in *Kluyveromyces fragilis* and *Saccharomyces fragilis*.^{66,}
⁶⁷ As we do not expect computer docking to be able to distinguish between an agonist and an antagonist, the effective hit rate of the primary screen was therefore 58% at 100 μ M.

The eleven 100- μ M inhibitors were subsequently tested at 10 μ M. Three showed > ~50% average inhibition; compounds **1**, **2**, and **3** (clorobiocin) had IC₅₀ values of 3.6 ± 0.8 μ M, 5.6 ± 0.8 μ M, and 5.0 ± 1.2 μ M, respectively, and Hill slopes of 1.8 ± 0.6 , 1.2 ± 0.3 , and 2.1 ± 0.5 , respectively (Tables 1). Interestingly, compounds **1** and **2** share a 2'-(phenylcarbamoyl)-[1,1'-biphenyl]-2-carboxylic acid core scaffold.

In one recent study, 95% of the inhibitors identified in a high-throughput screen acted through a nonspecific aggregation-based mechanism. This same study suggested that aggregation-based inhibition typically produces steep Hill slopes that are much greater than unity, with average values around 2.2.⁶⁸ As the Hill slopes of compounds **2** and **3** (clorobiocin) were significantly greater than unity, *TbGalE*

inhibition was measured at 30 μ M inhibitor concentration ($\sim 5 \times \text{IC}_{50}$) in the presence and absence of a detergent that disrupts colloidal aggregates (0.06% n-octylglucopyranoside). No significant differences in inhibition were noted, suggesting that inhibition is specific rather than aggregation-based.³¹ Given that the possibility of aggregation was eliminated, the steep Hill slopes of these two compounds provide further evidence for allostery between the two monomers of the *TbGalE* dimer.^{69, 70}

Clorobiocin: An Interesting Inhibitor

One of the hits from the primary virtual screen, compound **3** (clorobiocin), an aminocoumarin derived from several *Streptomyces* species, has previously been shown to inhibit the growth of *Trypanosoma cruzi* (*T. cruzi*), a close relative of *T. brucei*.⁷¹ As clorobiocin is a known bacterial topoisomerase II inhibitor, some have hypothesized that topoisomerase II may be the *T. cruzi* protein target as well,⁷² though other targets could not be ruled out.⁷¹ The current work suggests that glucose 4'-epimerase may also be among the proteins targeted by this apparently polypharmacophoric compound.

We note with interest that novobiocin, a compound structurally similar to clorobiocin that likewise inhibits the growth of *T. cruzi*,⁷¹ did not show greater than 50% *TbGalE* inhibition at 10 μ M, despite the fact that our computational model predicted high binding affinity. The crude scoring functions employed by AutoDock and Vina, optimized not only for accuracy but also for speed, seem unable to

differentiate between the apparently subtle differences in the protein-ligand interactions of clorobiocin and novobiocin.

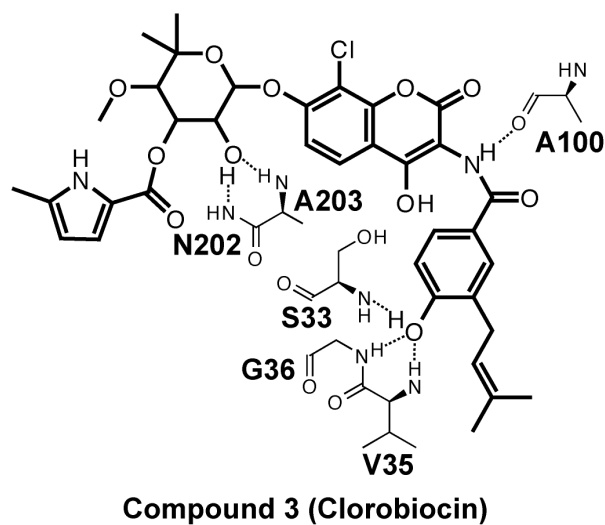


Figure 1: The predicted ligand-protein interactions of clorobiocin bound to *TbGalE*, shown schematically.

Predicted binding poses of top inhibitors

Computer docking suggests that compound **3** (clorobiocin) occupies the NAD^+ -binding pocket. To further characterize the clorobiocin binding pose, we examined the Vina scores of clorobiocin docked into each of the protein configurations of the ensemble and selected the binding pose / protein configuration associated with the best score for further analysis. The predicted protein-ligand interactions are represented schematically in Figure 1.

Table 1: The fourteen low-micromolar *Tb*GaIE inhibitors identified. The first column shows compound structures, and the second column shows the compound identification number used throughout the text. The third column shows the IC₅₀ value of each compound as measured in enzymatic assays. The fourth and fifth columns show the EC₅₀ values as measured in whole-cell assays of *T. brucei* and human MRC5 cells, respectively. The final column shows the predicted LogP value of each compound.

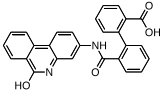
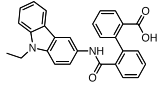
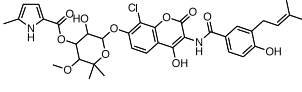
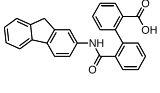
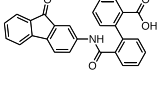
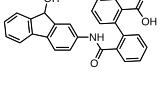
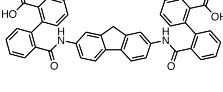
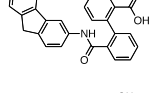
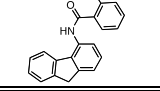
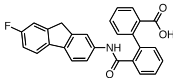
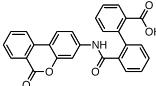
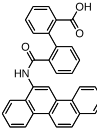
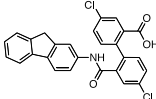
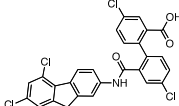
Structure	ID	TbGaIE IC ₅₀ (μM)	Tb EC ₅₀ (μM)	MRC5 EC ₅₀ (μM)	LogP
	1	3.6 ± 0.8	> 50	> 50	4.74
	2	5.6 ± 0.8	> 50	> 50	6.16
	3	5.0 ± 1.2	= 4.4	> 50	1.12
	4	4.6 ± 1.4	> 50	> 50	5.62
	5	2.8 ± 0.7	> 50	> 50	4.47
	6	6.8 ± 6.8	> 50	> 50	4.57
	7	0.9 ± 0.05	> 50	> 50	7.40
	8	4.2 ± 1.2	> 50	> 50	5.62
	9	8.0 ± 3.3	> 50	> 50	5.49

Table 1 continued:

Structure	ID	TbGalE IC ₅₀ (μM)	Tb EC ₅₀ (μM)	MRC5 EC ₅₀ (μM)	LogP
	10	3.8 ± 0.8	> 50	> 50	5.92
	11	3.2 ± 0.8	> 50	> 50	4.29
	12	2.6 ± 0.4	24.4	> 50	6.44
	13	3.3 ± 1.2	28.5	> 50	6.60
	14	1.8 ± 0.8	> 50	35.2	7.48

Clorobiocin is predicted to participate in multiple hydrogen bonds with the backbone atoms of amino acids lining the NAD⁺ binding pocket. One ligand hydroxyl group is predicted to form two hydrogen bonds with the backbone amines of N202 and A203; a second ligand hydroxyl group is predicted to form hydrogen bonds with the backbone amines of V35, G36, and S33. Finally, a ligand secondary amine is predicted to form a hydrogen bond with the backbone carbonyl oxygen atom of A100. We note again, however, that many of these same protein-ligand interactions characterize the predicted binding mode of novobiocin as well (data not shown), despite the fact that novobiocin binding to *TbGalE* is weak.

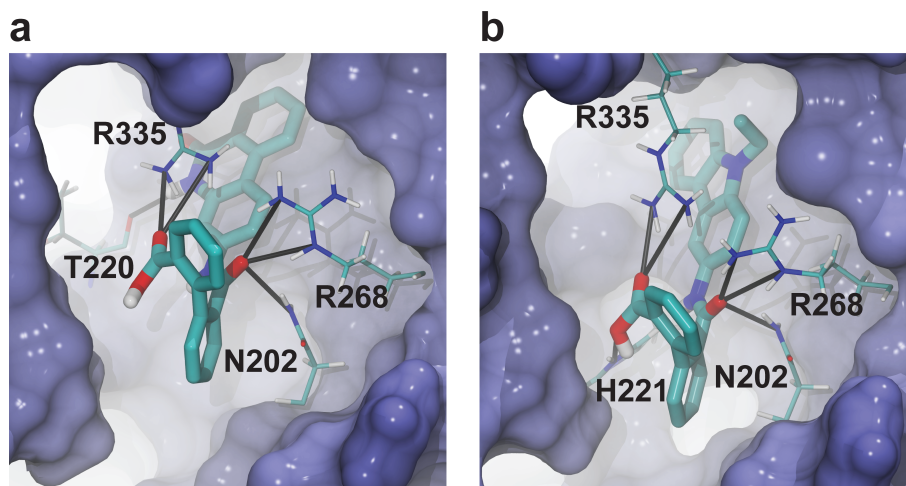


Figure 2: The predicted binding modes of compounds **1** and **2**. Some portions of the protein have been removed to facilitate visualization. Hydrogen bonds between the protein and ligand are shown as black lines.

The binding modes of compounds **1** and **2**, both predicted to bind in the UDP pocket, were not so paradoxical. These compounds contain similar 2'-(phenylcarbamoyl)-[1,1'-biphenyl]-2-carboxylic acid core scaffolds (Table 1) and similar predicted binding modes (Figure 2). For each of these two ligands, we again examined the Vina scores of the ligand docked into the various protein configurations of the ensemble and selected the binding pose / protein configuration associated with the best score for further analysis (Figure 2). The core scaffold common to compounds **1** and **2** participates in many of the same hydrogen bonds that characterize UDP-glucose binding. R335, N202, and R268 all form hydrogen bonds with the diphosphate moiety of UDP-glucose; they likewise form hydrogen bonds with the carboxylate group and the carbonyl oxygen atom of the core scaffold. Additionally, H221 may also form a hydrogen bond with the carboxylate group of the core scaffold (Figure 2).

Cation- π interactions also seem to play a critical role in the binding of compounds **1** and **2**. In these interactions, the positive charge of the cation is electrostatically attracted to the quadrupole moment of an aromatic group.⁷³ If the predicted binding of the core scaffold of compounds **1** and **2** is correct, both R335 and R268 participate in cation- π interactions with the ligand.

Compounds **1** and **2** have polyaromatic moieties that extend into the pocket normally occupied by the uridine and ribose moieties of UDP-glucose. Much work can yet be done to optimize these fused ring systems in order to improve binding, as hinted at by the hydrogen bond predicted to form between compound **1** and the backbone carbonyl oxygen of T220 (Figure 2a). Other potentially interacting groups in the uridine and ribose portions of the UDP-glucose binding pocket include the backbone carbonyl oxygen atom of P253 and the backbone amine of F255. It may also be possible to add moieties to the fused ring system that exploit the F255 aromatic side chain, which can participate in π - π and cation- π interactions.

Finally, the nucleophilic side-chain thiol of C266 is attractive from a drug-design perspective; fragments with electronegative moieties could be added to the fused ring systems of compounds **1** and **2** in order to facilitate the formation of a covalent adduct with the C266 thiol group, a strategy similar to that employed by Kerr et al. in designing potent vinyl-sulfone inhibitors of parasitic cysteine proteases, including K777, which, pending FDA approval, will soon enter phase I clinical trials.⁷⁴ Additionally, as this cysteine is a glycine in *HsGalE*, compounds that target C266 may therefore be selective for the trypanosomal form of the enzyme, though

caution is advised, as non-specific reactions with protein thiols could lead to toxic side effects.

The Secondary Relaxed Complex Screen: Similarity Search

Encouraged by these initial results, we next searched the entire NCI database for compounds similar to the three low-micromolar inhibitors verified experimentally. Eighty additional compounds were subjected to the same relaxed complex protocol used in the primary virtual screen, and 14 novel compounds were thus identified as potential binders. One of these compounds, kedarcidin, a structural orthologue of clorobiocin, was unavailable from the NCI, and another failed the purity criteria. Thus, 12 compounds were subsequently tested experimentally.

Eleven of the twelve compounds showed > 50% average inhibition at 10 μ M. Subsequent experimental analysis confirmed that these compounds had IC₅₀ values between 0.8 and 6.8 μ M (Table 1). As previously, aggregation effects were excluded, and compound identity and purity were confirmed by accurate mass and LC-MS.

Taken together with the compounds of the initial screen, these inhibitors constitute a novel hit series based on a 2'-(phenylcarbamoyl)-[1,1'-biphenyl]-2-carboxylic acid scaffold. Additionally, a singleton hit (clorobiocin) was also found to be potent, though novobiocin, a related compound that might have otherwise been considered a member of the same hit series, was not a potent *Tb*GalE inhibitor.

Whole-cell Assays

All low-micromolar *TbGale* inhibitors identified were tested for their ability to inhibit the growth of cultured *T. brucei* and human liver MRC5 cells using the established Alamar Blue protocol.^{62, 63} Two compounds containing the 2'-(phenylcarbamoyl)-[1,1'-biphenyl]-2-carboxylic acid core scaffold, compounds **12** and **13**, had EC₅₀ values of 24.4 and 28.5 μ M against whole-cell *T. brucei*, respectively. Additionally, the natural product clorobiocin (compound **3**) had an EC₅₀ value of 4.4 μ M. Only compound **14**, a compound with no activity against whole-cell *T. brucei*, demonstrated any inhibition of MRC5 growth (Table 1).

To better understand why most of the *TbGale* inhibitors failed to inhibit whole-cell *T. brucei*, we calculated the LogP value of each (Table 1). With only one exception, the logP values were all high, either near the upper bound for what is considered "druglike" or, in several cases, well beyond that bound.^{75, 76} As these compounds are very hydrophobic, we postulate that they are retained in the cellular membrane, explaining the reduced efficacy against whole-cell *T. brucei*. The one compound with a druglike LogP value was compound **3** (clorobiocin), which has a measured whole-cell EC₅₀ of 4.4 μ M, suggesting inhibition of intracellular targets including *TbGale* and, potentially, topoisomerase II.⁷² Alternatively, the charged carboxylate moiety characteristic of the main hit series, absent in clorobiocin, may likewise prevent compound traversal of the cellular membrane.

Designing compounds similar to those based on the 2'-(phenylcarbamoyl)-[1,1'-biphenyl]-2-carboxylic acid core scaffold but with reduced hydrophobicity

should not be difficult. We note that these inhibitors have roughly the same IC_{50} values regardless of the fused ring system that is attached to the core scaffold. This suggests that the fused ring systems likely contribute to the overall binding affinity principally through non-specific interactions, if at all. Hydrophilic functional groups could likely be added to the fused ring systems without compromising the overall binding affinity, thereby increasing the propensity of these molecules to traverse cellular membranes.

Several of the fused ring systems have functional groups that may be amenable to chemical modification. Hydrophilic moieties could be added to the fused ring systems of compounds **1** and **6** *via* substitution at the hydroxyl group, for example. Several of the *TbGalE* inhibitors identified also have aromatic halides that could facilitate fragment addition.

As these compounds are already quite large, however, a better approach may be to entirely substitute the fused ring systems of the compounds tested here with other more hydrophilic molecular fragments. Reactions between the readily available compound [1,1'-biphenyl]-2,2'-dicarboxylic acid (NSC1966) and varied aromatic and even non-aromatic amines could expand the chemical series explored here to include compounds with more favorable partition coefficients.

If the above modifications fail to produce high-affinity inhibitors of cellular growth, the charged carboxylate of the main-series compounds could be methylated to potentially facilitate membrane traversal. Once in the cell, this prodrug could be easily demethylated by cellular esterases.⁷⁷

A Comparison of the Relaxed Complex Screen versus a Traditional Virtual Screen

Though not used for inhibitor identification, four traditional virtual screens of the NCI Diversity Set II against a *TbGalE* crystal structure were performed using Vina for the sole purpose of assessing how accounting for receptor flexibility altered compound ranking. Like the relaxed complex screens used to identify *TbGalE* inhibitors, the four traditional screens differed only in the binding pocket that was targeted (UDP-glucose binding pocket, NAD⁺ cofactor present; UDP-glucose pocket, NAD⁺ pocket, or combined UDP-glucose / NAD⁺ pockets, NAD⁺ cofactor absent). The best rank that compounds **1**, **2**, and **3** (clorobiocin) obtained in any of these four screens was 202nd, 10th, and 1st, respectively. Thus, while compounds **1** and **3** would have likely been identified even if receptor flexibility had been ignored, compound **2** would not have been identified in a traditional virtual screen against the static crystal structure alone.

Methodological Comparison

The primary relaxed complex screen was specifically designed in order to facilitate a comparison of different relaxed complex methodologies. While the comparisons made here may be system specific, we believe they may nevertheless provide useful insights for future studies.

First, did rescoring the Vina-docked poses with the AutoDock 4.0 scoring function help identify inhibitors that would have been otherwise missed? None of the five additional compounds identified by AutoDock rescoring exhibited > 50%

inhibition of *TbGalE* at 10 μ M; one, compound **S7**, exhibited > 50 % inhibition at 100 μ M, and one, compound **S16**, exhibited > 2 fold induction at 100 μ M. As our goal was to identify low-micromolar *TbGalE* inhibitors, rescoring with the AutoDock scoring function provided little advantage over Vina scoring alone.

Second, which ensemble-based ranking criteria (ensemble-average vs. ensemble-best) was best at identifying true positives? Consider the 26 primary-screen compounds identified using the Vina scoring function. Had we ranked the compounds by the ensemble-average score alone, seven experimentally tested compounds would have been discarded, including four 100- μ M inhibitors, one 10- μ M inhibitor, and one 100- μ M agonist. Had we ranked the compounds by the ensemble-best score alone, six experimentally tested compounds would have been discarded, including one 10- μ M inhibitor and two 100- μ M agonists. Thus, both ensemble-based ranking criteria identified true binders that the other missed. We recommend using both ranking criteria in future relaxed complex screens.

2.4 Conclusion

In the current study, we used computer-aided drug design to identify fourteen inhibitors of *TbGalE*, a potential *T. brucei* drug target. As novel HAT therapeutics are urgently needed, we are hopeful that the hit series described here will serve as a useful scaffold for further drug optimization. Our study also demonstrates the importance of accounting for receptor flexibility when predicting small-molecule protein inhibition,

as one of the primary-screen inhibitors would not have been identified had we conducted a virtual screen against the crystal structure alone.

This study also has implications for the application of the relaxed complex scheme.²⁸ In our lab's traditional implementation of the scheme, compounds have been ranked by the ensemble-average docking score. However, we here show that the ensemble-best score is also a useful ranking metric.

The chemical series of *Tb*GaleE inhibitors based on a 2'-carbamoyl-[1,1'-biphenyl]-2-carboxylate core scaffold is promising. Future efforts should focus on improving target and cellular potency. As the fused ring systems seem to contribute to ligand binding only nonspecifically, if at all, they could potential be replaced. The side-chain thiol of C266, which comes within 4 Å of the predicted binding mode of compound **2**, could also be exploited for more specific binding. As this cysteine is absent from the active site of the corresponding human protein, inhibitors that exploit C266 may also be more pathogen specific.

2.5 Acknowledgements section

JDD is funded by a Pharmacology Training Grant through the UCSD School of Medicine. MDU and MAJF are funded by a Programme Grant from The Wellcome Trust (085622). This work was also carried out with funding from NIH GM31749, NSF MCB-0506593, and MCA93S013 to JAM. Additional support from the Howard Hughes Medical Institute, the National Center for Supercomputing Applications, the San Diego Supercomputer Center, the W.M. Keck Foundation, the National

Biomedical Computational Resource, and the Center for Theoretical Biological Physics is gratefully acknowledged. All compounds were provided by the NCI/DTP Open Chemical Repository (<http://dtp.cancer.gov>). We would also like to thank Matthew Durrant for help with virtual screening.

Chapter 2, in full, is a near reprint of material that was recently accepted for publication in the Journal of Medicinal Chemistry. The dissertation author and Michael D. Urbaniak were the primary co-contributors to this work. Michael A. J. Ferguson and J. A. McCammon served as the advisers and principal investigators. The Journal of Medicinal Chemistry does not consider academic theses to constitute prior publication and so implicitly authorizes the inclusion of this chapter.

CHAPTER 3: Novel Naphthalene-Based Inhibitors of *Trypanosoma Brucei* RNA Editing Ligase I

Jacob D. Durrant, Laurence Hall, Achim Schnaufer, Rommie E. Amaro

3.1 Introduction

Trypanosoma brucei (*T. brucei*) is the causative agent of Human African trypanosomiasis (HAT), also known as African sleeping sickness. Neglected tropical diseases, including illnesses caused by trypanosomes like HAT, Chagas disease, and Leishmaniasis, are responsible for the loss of an estimated 56.6 million disability-adjusted life-years across several regions, particularly Sub-Saharan Africa.⁷⁸ Preventative measures such as vector control are effective at decreasing the incidence of HAT; however, given infection, the current treatments are unacceptable on the whole, particularly once *T. brucei* has infiltrated the central nervous system.⁷⁹

First-stage treatments include pentamidine and suramin, drugs developed more than half a century ago. Unfortunately, these drugs have severe side effects. Pentamidine is associated with hypoglycaemia and hypotension, and suramin is associated with anaphylactic shock, neurotoxic signs, severe cutaneous reactions, and renal failure.⁸⁰ The most common treatment for second-stage HAT is melarsoprol, a highly toxic drug with a 3% - 10% fatality rate.⁸¹ The danger of treatment is only compounded as more melarsoprol-resistant parasites emerge, particularly in central Africa.⁸² Eflornithine, another HAT treatment, is less toxic but is only effective against the *T. b. gambiense* subspecies; additionally, eflornithine is more costly to produce than melarsoprol.⁴³ Given the weaknesses of current treatments, new drugs are urgently needed.

Fortunately, recent studies of the trypanosomal editosome have revealed several new drug targets. In trypanosomes, the central dogma of biology includes an extra RNA-editing step. As in other eukaryotes, mitochondrial DNA is transcribed into RNA. In *T. brucei*, however, a protein complex known as the editosome makes extensive uridylyte insertions and deletions following transcription, at times even doubling the length of the original RNA sequence.^{83, 84} Following the editing process, a nick in the RNA remains; RNA editing ligase 1 (*TbREL1*) is responsible for religation. As the editosome is absent in humans, the proteins of this complex, including *TbREL1*, are potential drug targets.

Recently, Amaro et al. identified several novel *TbREL1* inhibitors. The relaxed complex scheme, a virtual-screening methodology that accounts for full protein flexibility,²⁸ was used to identify five low-nanomolar inhibitors from among the compounds of the National Cancer Institute Diversity Set I.³¹ Unfortunately, these *TbREL1* inhibitors were ineffective in whole-cell assays, perhaps in part because they are too hydrophilic to cross lipid membranes (unpublished work).

Motivated by both the urgent need for novel trypanocidal therapeutics as well as the success of virtual screening against *TbREL1* in the past, we here use the relaxed complex scheme to indentify additional naphthalene-based inhibitors in hopes of finding compounds that can kill whole-cell *T. brucei*. First, online databases of commercially available compounds were searched for compounds similar to the inhibitors previously characterized. Following virtual screening, the most promising of these compounds were subsequently tested experimentally, revealing four novel

*Tb*REL1 inhibitors with unique naphthalene-based scaffolds, two of which have ALogP values that suggest reasonable lipophilicity.

3.2 Materials and Methods

Online Similarity Search

To generate a library of compounds similar to the *Tb*REL1 inhibitors characterized previously,³¹ we performed online substructure searches of several databases of commercially available compounds, including Hit2Lead (Hit2Lead.com, ChemBridge), the NCI/DTP Open Chemical Repository (dtp.cancer.gov), Sigma-Aldrich (sigmaaldrich.com), and ZINC.⁸⁵ Searches were performed using three structures similar to the core naphthalene scaffolds of known inhibitors: naphthalene-2-sulfonic acid, 2-naphthoic acid, and 2-nitronaphthalene (Figure 3).

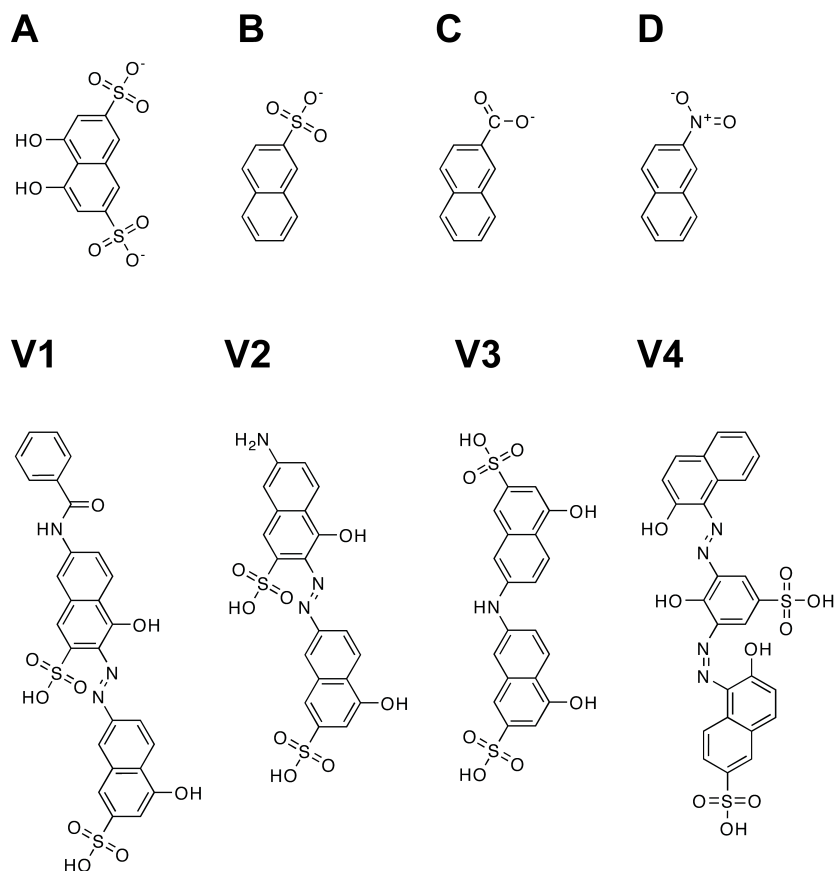


Figure 3: A) The core 4,5-dihydroxynaphthalene-2,7-disulfonate scaffold of three previously identified *TbREL1* inhibitors. B) Naphthalene-2-sulfonic acid. C) 2-naphthoic acid. D) 2-nitronaphthalene. V1-V4) The four low micromolar inhibitors identified in the current work.

Initial Docking against the Crystal Structure

The compounds identified *via* online substructure searches were each docked into a 1.20-Å resolution crystal structure of *TbREL1* (PDB ID: 1XDN⁸⁶) using AutoDock 4.²⁶ Ligand files were processed with AutoDockTools 1.4.5 to merge nonpolar hydrogens with parent heteroatoms and to assign gasteier charges. AutoGrid affinity grids contained 86 x 72 x 78 points spaced 0.375 Å apart, centered on the *TbREL1* active site. Grid files were created for the following ligand atom types: A

(aromatic carbon), C, F, I, N, NA (hydrogen-bond accepting N), Cl, OA (hydrogen-bond accepting O), P, S, SA (hydrogen-bond accepting S), Br, HD (hydrogen-bond hydrogen), and e (electrostatic).

AutoDock parameters similar to those published previously by Amaro et al.³¹ were used: population size 200; 5,000,000 evaluations; 27,000 generations; 100 runs; and cluster tolerance of 2.0 Å. All other AutoDock parameters were set to the default values. The correct docked pose was judged to be the lowest-energy pose of the most populated cluster.

MD Simulation

With the intent of rescoring the top hits from the initial crystal-structure screen in a way that accounts for full protein flexibility, we drew upon a previous study of *TbREL1* molecular motions.⁸⁷ In brief, molecular dynamics (MD) simulations of *TbREL1* were performed using NAMD 2.6.⁵⁹ Four hundred receptor conformations were extracted from the MD simulations, one every 50 ps. QR factorization⁸⁸ was used to eliminate conformational redundancy, thereby reducing the number of representative structures from 400 to 33.³¹ These 33 *TbREL1* structures are said to constitute an *ensemble* representative of the many protein conformations sampled during the MD simulation.

Relaxed Complex Screen

The relaxed complex scheme (RCS) was subsequently used to rescore the top compounds from the initial crystal-structure screen.²⁸ AutoDock was used to dock each of the top inhibitors into the 33 protein configurations of the receptor ensemble using the same docking parameters described above. The ensemble-average binding energy of each ligand was computed by taking the simple mean, and the ligands with the best mean predicted binding energy were subsequently tested experimentally.

Experimental Validation

The top compounds from the secondary relaxed complex screen were ordered for use in an experimental assay. The Developmental Therapeutic Program at the National Cancer Institutes of Health provided V1, V2, and V3 (Figure 3), which were dissolved in DMSO. Unless otherwise noted, all other reagents were purchased from Sigma.

The protocols for recombinant *TbREL1* expression, purification, and assaying have been described previously.³¹ In brief, the *TbREL1* coding sequence was cloned into a pFastBac1 plasmid, which was subsequently transformed into DH10Bac *Escherichia coli*. Recombinant *TbREL1* was then purified from transfected Sf9 insect cells via a C-terminal tandem affinity purification (TAP) tag.

To measure enzyme inhibition, *TbREL1* was incubated with [³²P]AMP and varying concentrations of the potential inhibitors. The extent of protein adenylation was subsequently measured by SDS/PAGE and phosphorimaging. All reactions were

done in at least triplicate, and IC₅₀ values were calculated using the GraphPad Prism 5 software.

3.3 Results/Discussion

T. brucei RNA editing ligase 1 (*TbREL1*) is a component of the trypanosomal editosome. In trypanosomes, mitochondrial RNA is edited following transcription; after the editing process is complete, *TbREL1* religates the RNA in preparation for ribosomal translation. *TbREL1* is a viable drug target because it is required for *T. brucei* survival^{89, 90} and because it has no close human homologues.³¹ The heavy disease burden caused by Human African trypanosomiasis (HAT), the urgent need for novel trypanocidal therapeutics, and the success of virtual screening against *TbREL1* in the past have motivated the current work, wherein we identify novel *TbREL1* inhibitors with naphthalene-based scaffolds.

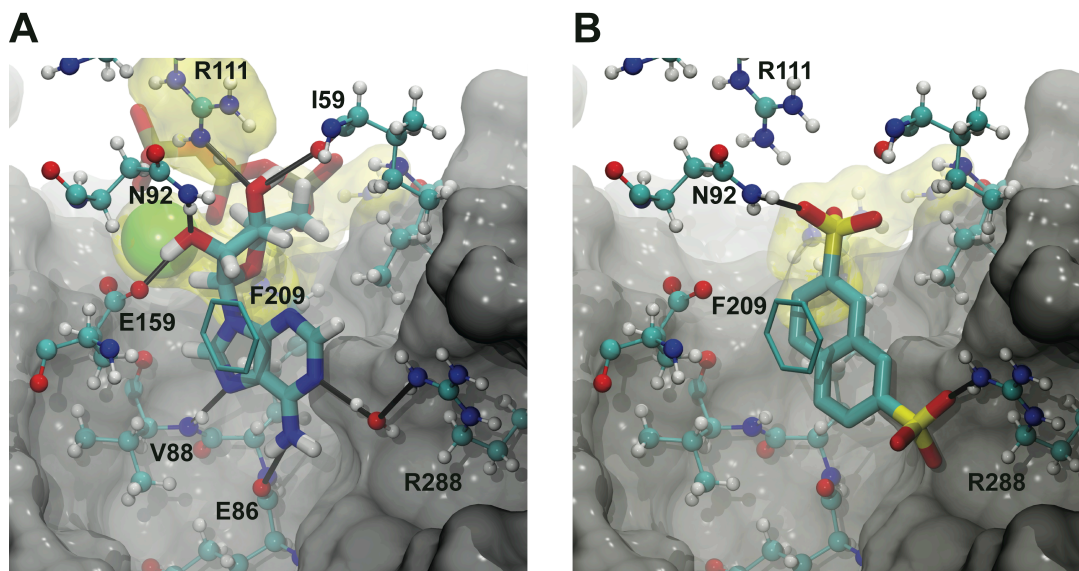


Figure 4: *TbREL1* binding. Solid black lines represent hydrogen bonds. Electropositive residues and ions (Mg²⁺) at the active-site periphery are highlighted in yellow. The carbons of the F209 phenyl ring are shown in licorice. Portions of the protein are absent to improve clarity. A) ATP binding. B) The predicted binding pose of naphthalene-2,7-disulfonate.

Similarity Search

Previously, Amaro et. al identified several micro-molar inhibitors of *TbREL1*.³¹ The top three inhibitors identified were all based on a naphthalene-2,7-disulfonate (NDS) scaffold. Computer docking provides insight into why this scaffold is amenable to *TbREL1* inhibition (Figure 4). Similar to the adenine moiety of ATP (the native substrate), the NDS naphthalene group is able to form π - π stacking interactions with F209. Additionally, one of the negatively charged NDS sulfonate groups interacts with R111 at the active-site periphery; this same protein residue also participates in electrostatic and hydrogen-bond interactions with the ATP polyphosphate tail. A hydrogen bond is also formed between NDS and N92, similar to the hydrogen bond formed with the O2' oxygen atom of the ATP ribose. Finally,

docking suggests that the second of the two NDS sulfonate groups is buried deep within the binding pocket, displacing a water molecule that normally mediates a hydrogen-bond network between the ATP adenine N1 atom and R288. This water displacement allows the sulfonate group to interact with the charged R288 residue directly.

Unfortunately, these previously identified *Tb*REL1 inhibitors were ineffective in whole-cell assays, likely because they are too hydrophilic to cross lipid membranes (unpublished work). In an attempt to identify additional naphthalene-based *Tb*REL1 inhibitors that may fare better, we searched several online databases of commercially available compounds for similar structures: naphthalene-2-sulfonic acid, 2-naphthoic acid, and 2-nitronaphthalene (Figure 3). These searches identified 588 compounds: 61 compounds from Hit2Lead (Hit2Lead.com, ChemBridge), 394 from the NCI/DTP Open Chemical Repository (dtp.cancer.gov), 87 from Sigma-Aldrich (sigmaaldrich.com), and 46 from ZINC.⁸⁵ In all, the search identified 376 naphthalene-2-sulfonic acid compounds, 130 2-naphthoic acid compounds, and 85 2-nitronaphthalene compounds.

Virtual Screen

AutoDock 4.0 was used for docking because it has been successfully used to identify *Tb*REL1 inhibitors in the past.³¹ The AutoDock scoring function is much faster, though less accurate, than more rigorous methodologies such as thermodynamic integration,^{26, 91} single-step perturbation,⁹² and free energy perturbation.⁹³

Additionally, AutoDock performs well⁹⁴ when compared to other docking programs, including DOCK,⁹⁵ Flex,⁹⁶ and GOLD.⁹⁷

The 588 compounds identified through online substructure searches were first docked into a 1.20-Å resolution crystal structure of *TbREL1*.⁸⁶ AutoDock placed 14% of the naphthalene compounds in the expected pose (26% of the 2-naphthoic acid compounds, 10% of the naphthalene-2-sulfonic acid compounds, and 8% of the 2-nitronaphthalene compounds), with the naphthalene portion of the ligand buried deep in the pocket and the electronegative group at the two position either interacting with R288 or with R111 at the active-site periphery.

The preliminary docking to the *TbREL1* crystal structure, while useful for eliminating those structures that were grossly incompatible with the *TbREL1* active site, did not account for full protein flexibility. Aside from the inaccuracies inherent in docking scoring functions themselves, docking suffers from additional inaccuracies when protein flexibility is ignored. When a ligand approaches a protein receptor *in vitro*, it does not encounter a single static structure, but rather an ensemble of many different protein configurations. Often, a given ligand may only bind a certain subset of all protein configurations sampled, depending in part on the varied side-chain positions of active-site residues. When multiple protein conformations are incorporated into a virtual-screening protocol, the hit rate can drastically improve; ligands that do not bind to the crystal structure may bind to other related protein configurations. Screening against these other configurations in principle reduces the false negative rate.

In order to account for full protein receptor flexibility, the top 45 compounds from the initial crystal-structure screen, roughly corresponding to the top 7.5% of the library, were docked into 33 protein receptor configurations extracted from a MD simulation of *TbREL1*.³¹ The 45 ligands were then reranked by their respective ensemble-average scores, and the top twelve compounds were subsequently tested experimentally.

Table 2: IC_{50} is a measure of inhibition, in μM ; $AutoDock_{Crystal}$ is the predicted binding energy to the crystal structure; $Rank_{Crystal}$ is the rank of the ligand when all 588 compounds are ordered by their respective $AutoDock_{Crystal}$ values; $AutoDock_{Ensemble}$ is the average predicted binding energy to 33 representative protein-receptor configurations; $Rank_{Ensemble}$ is the rank of the ligand when the top 45 compounds are ordered by their respective $AutoDock_{Ensemble}$ values; % Expected Pose is the percentage of the 33 representative protein structures amenable to deep-pocket binding, in which the naphthalene core is docked deep into the binding pocket; $AutoDock_{Ensemble/Expected}$ is the average predicted binding energy when only those members of the ensemble amenable to deep-pocket binding are considered; $AutoDock_{Ensemble/Unexpected}$ is the average predicted binding energy when only the remaining members of the ensemble are considered; and $AutoDock_{Ensemble/Best}$ is the predicted binding energy of the ligand to the “optimal protein configuration” from the ensemble. All predicted energies are in kcal/mol.

	V1	V2	V3	V4
IC_{50}	2.16 ± 1.20	1.53 ± 1.17	8.36 ± 1.71	1.59 ± 1.1
$AutoDock_{Crystal}$	-11.8	-11.3	-10.7	-11.8
$Rank_{Crystal}$	11	20	31	25
$AutoDock_{Ensemble}$	-11.9 ± 1.4	-11.9 ± 1.4	-10.2 ± 1.0	-12.8 ± 1.6
$Rank_{Ensemble}$	4	3	12	1
% Expected Pose	33%	33%	24%	18%
$AutoDock_{Ensemble/Expected}$	-13.3 ± 1.2	-12.6 ± 1.1	-11.6 ± 0.5	-14.9 ± 0.6
$AutoDock_{Ensemble/Unexpected}$	-11.2 ± 0.9	-11.6 ± 1.4	-9.8 ± 0.6	-12.3 ± 1.4
$AutoDock_{Ensemble/Best}$	-15.3	-13.9	-12.1	-15.6

Experimental Results

Prior to RNA ligation, a key *TbREL1* lysine must first be adenylylated. To measure the inhibition of this first step of the reaction pathway, the formation of *TbREL1*-[³²P]AMP was monitored *via* SDS/PAGE and autoradiography in the

presence of predicted inhibitor. Triton X-100 (0.1%) was added in order to prevent inhibition by promiscuous, aggregate-based inhibitors. Four compounds, V1, (E)-7-benzamido-4-hydroxy-3-((5-hydroxy-7-sulfonaphthalen-2-yl)diazenyl)naphthalene-2-sulfonic acid; V2, (E)-7-amino-4-hydroxy-3-((5-hydroxy-7-sulfonaphthalen-2-yl)diazenyl)naphthalene-2-sulfonic acid; V3 (Di-J acid), and V4 (Mordant Black 25), inhibited *Tb*REL1 activity with IC_{50} values of $2.16 \pm 1.20 \mu\text{M}$, $1.53 \pm 1.17 \mu\text{M}$, $8.36 \pm 1.71 \mu\text{M}$, and $1.59 \pm 1.1 \mu\text{M}$, respectively (Table 2, Figure 3).

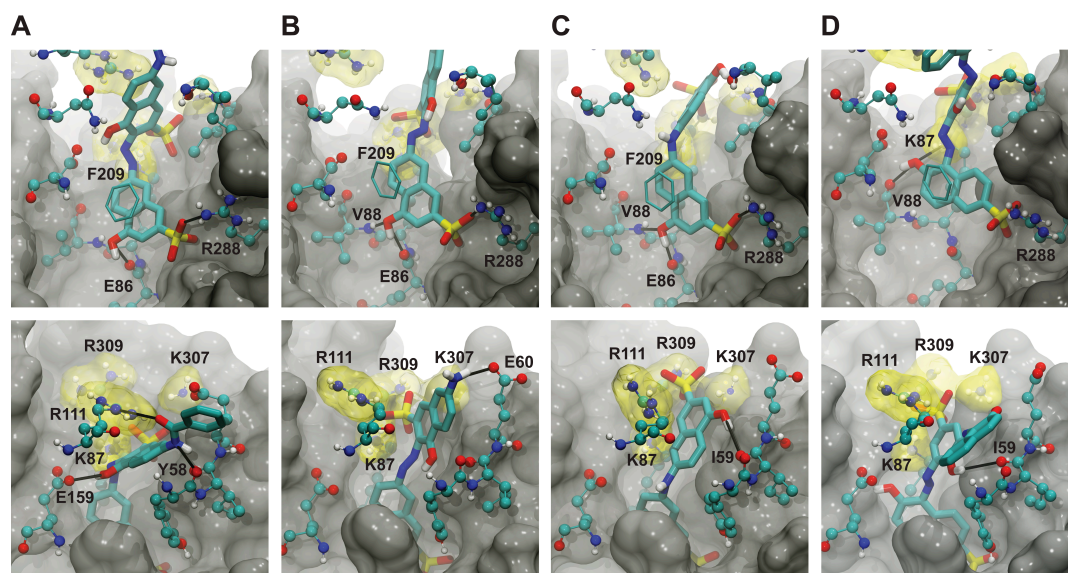


Figure 5: *Tb*REL1 binding. The top rows show binding deep within the active site, and the bottom rows show binding at the active-site periphery. Solid black lines represent hydrogen bonds. Electropositive residues at the active-site periphery are highlighted in yellow. The carbons of the F209 phenyl ring are shown in licorice. Portions of the protein were removed to improve clarity. A) The predicted binding pose of V1. B) The predicted binding pose of V2. C) The predicted binding pose of V3. D) The predicted binding pose of V4.

Analysis of Predicted Binding Poses

In order to analyze the predicted binding mode of the four confirmed *Tb*REL1 inhibitors, the protein configuration from the ensemble that gave the best AutoDock-predicted binding energy (the “optimal receptor”) was visualized together with the

associated docked ligand. In all cases, the electronegative group at the naphthalene two position was buried deep within the active site, forming interactions with R288, as expected. Additionally, three of the four ligands, like the three most potent *TbREL1* inhibitors identified previously,³¹ had hydroxyl groups in the naphthalene 4 position, suggesting that the hydrogen bonds formed with E86 and V88 are also critical to ligand binding (Figure 5, upper rows). A fourth ligand, V4, had a hydroxyl group in the naphthalene 6 position, where it could form hydrogen bonds with the backbone carbonyl oxygen atom of V88 and the side-chain amino group of K87.

At the active-site periphery, all four of the confirmed inhibitors had secondary sulfonate groups that docked near the more electropositive side of the active-site periphery, opposite the R111 residue. The peripheral electronegative sulfonate groups of previous NDS inhibitors, substituents of the naphthalene core itself, interacted principally with R111, but the peripheral sulfonate groups of the four inhibitors identified here are positioned on the more electropositive side of the active site, where they interact with K307, R309, and K87 (Figure 5, bottom rows). R111 is not entirely neglected, however; all four compounds are predicted to participate in π -cation interactions with R111.

In addition to these electrostatic interactions, the four inhibitors identified are also predicted to interact with other protein residues at the active-site periphery as well (Figure 5, bottom rows). In some ways, these interactions mimic the interactions between *TbREL1* and its native substrate, ATP. V1 forms a hydrogen bond with the R111 guanidinium group, similar to the bond formed between R111 and the ATP γ

phosphate. V1 also forms a hydrogen bond with the E159 side-chain carboxylate group, similar to the bond formed with the ATP 2' ribose hydroxyl group. In some ways, however, V1 binding to *TbREL1* is unique; V1 forms a hydrogen bond with the backbone carbonyl of Y58, a residue that does not participate in ATP binding (Figure 5A).

V2 is predicted to participate in only one hydrogen bond at the active-site periphery. This bond is formed with the E60 side-chain carboxylate group, a group that does not participate in ATP binding (Figure 5B). V3 and V4 are likewise predicted to form only one hydrogen bond at the active-site periphery, a bond with the I59 backbone carbonyl. This same backbone carbonyl forms a hydrogen bond with the 3' hydroxyl group of the ATP ribose (Figure 5c).

Ligand Binding may Occur via a Population-Shift Mechanism

For all four confirmed inhibitors, only one fifth to one third of the ensemble protein configurations permitted the expected docked pose, in which the naphthalene and sulfonate group of the ligand are buried deep in the binding pocket, as evidenced by the bimodal distribution of three of the four predicted binding energies to the various ensemble protein configurations (data not shown). The average predicted binding energy of the ligands was significantly higher when only those protein configurations amenable to deep-pocket binding were considered (Table 2). This result suggests that *in vitro* the four ligands identified may encounter many different *TbREL1* configurations as they approach the receptor, but that only a subset of those

configurations is amenable to tight binding. Upon binding, the population of protein configurations presumably shifts to accommodate the bound ligand. This population-shift model of ligand binding contrasts with the traditional “lock-and-key” model, in which the protein principally samples only those configurations that are amenable to binding without requiring a shift in population to accommodate the ligand.

Binding to the Crystal-Structure Protein Configuration may be Suboptimal

Interestingly, the crystal-structure protein configuration used for the initial docking is likely itself suboptimal for the binding of the four inhibitors identified, as evidenced by the improvement in rank when an ensemble-average AutoDock score was used ($\text{Rank}_{\text{Ensemble}}$) instead of the crystal-structure score ($\text{Rank}_{\text{Crystal}}$, Table 2). In fact, only one of the four compounds, V1, scored in the top twelve when all 588 compounds were docked into the crystal structure alone. V2, V3, and V4, which ranked 20th, 31st, and 25th against the crystal structure, respectively, may not have been tested had the ligand set not been reranked by an ensemble-average AutoDock score. A direct comparison of the predicted binding energy of the four identified inhibitors docked into the crystal structure ($\text{AutoDock}_{\text{Crystal}}$) and docked into the optimal protein configuration from the ensemble ($\text{AutoDock}_{\text{Ensemble/Best}}$) likewise demonstrates the importance of accounting for full protein flexibility; in all four cases, predicted energies of binding improved several kcal per mol when the optimal structure was used rather than the crystal structure (Table 2).

In order to investigate why binding to the crystal structure was suboptimal, the crystal structure was compared to the optimal receptor conformation for each of the four ligands. By aligning the optimal receptors to the crystal structure and visualizing both proteins and ligands, it is evident that in all four cases the crystallographic position of E60 prevented optimal binding. During the molecular dynamics simulation, however, E60 broke its contact with R111, opening up a new pocket that is occupied by all four of the novel inhibitors (Figure 6). This unique binding mode would not have been identified had protein-receptor flexibility been ignored.

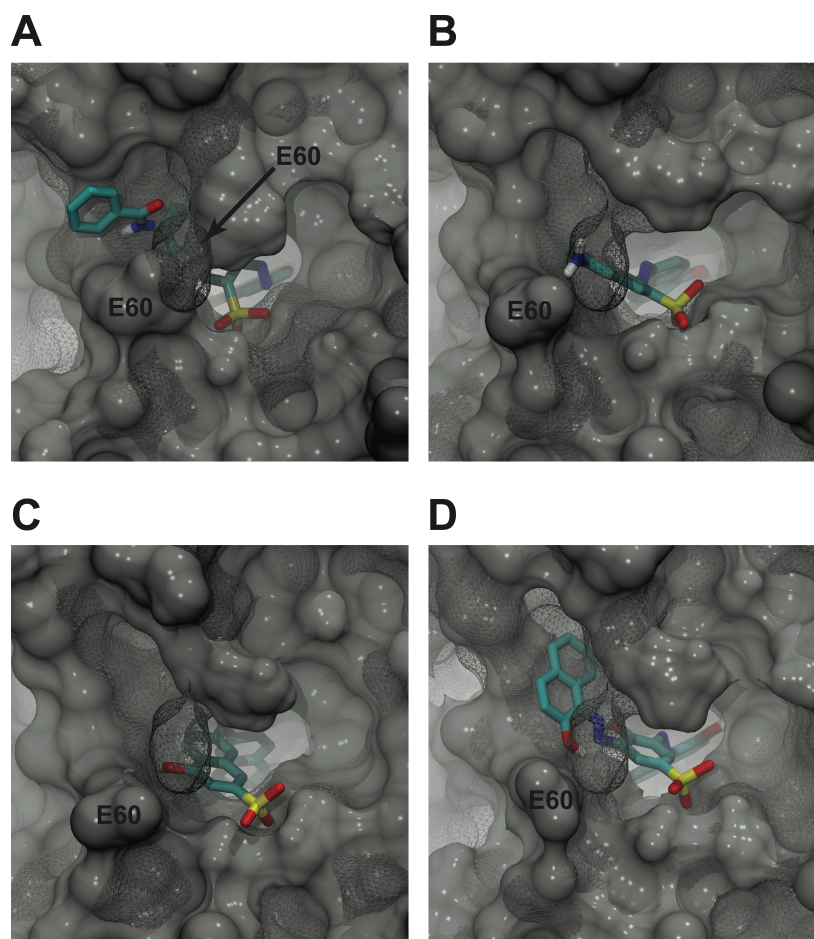


Figure 6: Predicted binding of the four identified inhibitors to the optimal protein configuration from the ensemble. The *TbREL1* crystal structure is shown in mesh. The E60 residues of the optimal configurations are labeled directly, and the E60 residue of the crystal structure is labeled with an arrow. In all four cases, the closed position of the crystal-structure E60 residue would have prevented optimal ligand binding. A) V1. B) V2. C) V3. D) V4.

3.4 Conclusion / Future Directions

Unfortunately, first-stage HAT treatments such as pentamidine and suramin have harsh side effects,⁸⁰ and second-stage treatments such as melarsoprol can be fatal. The pharmaceutical industry has been slow to develop novel trypanocidal therapeutics because HAT infections occur primarily in developing countries with little market appeal; indeed, the only novel trypanocidal therapeutic registered in the

last 50 years is eflornithine,⁹⁸ a drug that is likely only available because it can also be sold as a topical cream for the treatment of hirsutism in developed countries.

Given the hesitancy of the pharmaceutical industry, in recent years academics have played a leading role in the development of novel therapies to treat HAT. Amaro et al. recently demonstrated that a 4,5-dihydroxynaphthalene-2,7-disulfonate scaffold may be useful in the development of novel medicines that target *T. brucei* RNA editing ligase 1 (*TbREL1*), a known drug target. Unfortunately, these previously identified sulfonated naphthalene inhibitors, while effective inhibitors of the *TbREL1* protein, were ineffective in whole-cell assays. As Schrodinger's LigPrep software (Schrodinger) suggests that at pH 7.0 the sulfonates of these compounds are negatively charged, we hypothesize that they are too hydrophilic to cross the *T. brucei* lipid membrane in order to reach their physiological target. The ALogP's of Amaro's S5, V1, and S1 compounds were -1.043, -0.292, and -0.778, respectively (Discovery Studio, Accelrys), confirming that they are generally hydrophilic. Indeed, two of these three compounds, S5 and S1, are too hydrophilic to be considered druglike.⁷⁶

We have here built on the previous work of Amaro et al. to develop additional *TbREL1* inhibitors based on novel naphthalene scaffolds that may fare better in whole-cell assays. The compounds proposed in the current work are also sulfonated naphthalenes; however, some of them are more hydrophobic than the naphthalene-based inhibitors identified previously. The ALogP's of V1, V2, V3, and V4 are 0.492, -1.039, -1.112, and 1.835, respectively (Discovery Studio, Accelrys). Thus two of the novel inhibitors, V1 and V4, may actually prefer a lipid environment.

In the future, further drug optimization is needed. Three of the four novel compounds contain diazene linkers that may be hydrolysable *in vivo*. Furthermore, the nitrogen atoms of these linkers are not predicted to participate in hydrogen bonds with the protein; replacing one or both of them with carbon atoms may therefore decrease hydrophilicity without sacrificing compound potency. Additionally, some of the compounds contain other moieties like hydroxyl and amino groups that are not predicted to contribute to inhibitor binding. Perhaps these groups could likewise be eliminated.

Undoubtedly the charged sulfonate groups also contribute to the hydrophilicity of these compounds. In the virtual screen presented here, naphthalenes with carboxylic acids and nitro groups were included to see if the sulfonate groups could be replaced with less electronegative functional groups. Unfortunately, none of the compounds with carboxylate groups scored well enough to justify experimental testing, and the few nitro-group containing compounds that were tested failed to inhibit *TbREL1*. Rather than replacing the sulfonate groups, a better strategy may therefore be to modify those groups in order to neutralize their charge. For example, replacing the sulfonate groups with sulfonamides, a similar functional group that is not charged, may decrease hydrophilicity while preserving many important protein-ligand interactions.

3.5 Acknowledgements

JDD is funded by a Pharmaceutical Training Grant through the UCSD School of Medicine. Additional support from the Howard Hughes Medical Institute, the National Center for Supercomputing Applications, the San Diego Supercomputer Center, the W.M. Keck Foundation, Accelrys, Inc., the National Biomedical Computational Resource, the Center for Theoretical Biological Physics, and the laboratory of J. Andrew McCammon is gratefully acknowledged. REA acknowledges support from NIH F32-GM077729 and MRAC CHE060073N. AS is funded by NIH AI069057. Compounds V1, V2, and V3 were obtained from the NCI/DTP Open Chemical Repository and correspond to NSC45609, NSC162535, and NSC8674, respectively. We would also like to thank Matthew G. Durrant for help with writing and figure preparation, and Dr. Emmanuel Theodorakis for helpful discussions.

Chapter 3 is in large part a reprint of the material as it appears in a manuscript recently submitted to PLoS Neglected Tropical Diseases. The dissertation author and Laurence Hall were the primary co-contributors to this work. Achim Schnauffer and Rommie E. Amaro served as advisers and principal investigators. PLoS journals have an open-access policy that permits republication of PLoS articles in academic theses.

CHAPTER 4: Potential Drug-Like Inhibitors of Group 1 Influenza Neuraminidase Identified through Computer-Aided Drug Design

Jacob D. Durrant and J. Andrew McCammon

4.1 Introduction

Influenza is caused by RNA viruses of the family Orthomyxoviridae. While not generally life threatening in healthy adults, the virus occasionally mutates into more deadly forms and has been responsible for several pandemics in the last century. Recently, a new strain of pandemic influenza (H1N1) capable of infecting humans has been identified,⁹⁹ with a U.S. hospitalization rate of about 9%. Additionally, a distinct strain of avian influenza (H5N1) arose in 1997 that may cause a similar global pandemic in the future.¹⁰⁰

In preparation for pandemic influenza, many nations have stockpiled inhibitors of the influenza protein neuraminidase.¹⁰¹ Following formation, influenza viral particles remain bound to cell membranes via sialic-acid residues. Neuraminidase cleaves these residues, releasing the virus and enabling viral propagation.¹⁰² Neuraminidase is the target of several FDA-approved drugs, including zanamivir and oseltamivir,¹⁰¹ because it is essential for viral propagation and has a well-conserved active site.¹⁰³ Unfortunately, drug-resistant strains have recently emerged,¹⁰⁴⁻¹⁰⁷ and the need for novel inhibitors is great.

Motivated by the urgent need for new influenza therapeutics, we used AutoGrow,¹⁰⁸ a recently developed computer-aided drug-design program, to guide the design of several potential neuraminidase inhibitors predicted to bind better than currently approved drugs.

4.2. Material and Methods

Accounting for Protein Flexibility

To account for protein flexibility, we drew upon a molecular dynamics simulation of neuraminidase that has been described previously.¹⁰⁹ Protein conformations extracted from this 40-ns simulation were clustered into 27 groups by root-mean-square-deviation (RMSD) conformational clustering using the *gromos* clustering algorithm, as implemented in the GROMOS++ analysis software.^{61, 110} In brief, an RMSD distance was calculated for each pair of protein conformations extracted from the MD simulation. Those pairs with associated RMSD distances greater than 1.3 Å were discarded. The single conformation most frequently present in the remaining pairs, together with the other corresponding conformation of each pair, were merged into a list of conformations called the first cluster. The conformations of the first cluster were subsequently removed from the pool of conformations extracted from the MD simulation, and the process was repeated until no conformations remained. The centroid of each cluster was selected, producing an ensemble of 27 unique protein structures representative of the many protein conformations sampled during the simulation.

Initial AutoGrow Runs

AutoGrow was run three times, once using a neuraminidase crystal structure (PDB ID: 2HU4¹¹¹) as the template protein and twice using each of the top two

ensemble structures.¹⁰⁹ In each of these three runs, AutoGrow ran for eight generations, adding fragments to a core scaffold similar to oseltamivir. Each generation initially contained fifty ligands. For each generation after the first, ten primary individuals were taken from the previous generation, based on both the score of the most populated docking cluster and successful active-site binding. An additional twenty “crossovers” and twenty “mutants” were created from these ten primary individuals, subject to the requirement that all compounds contain fewer than one hundred atoms. The first generation initially contained only the scaffold and 49 “mutants,” as no previous generation existed from which “parents” could be drawn for crossover production.

To determine fitness, all AutoGrow-generated ligands were docked into their respective neuraminidase structures (the crystal structure or the two ensemble conformations) using AutoDock 4.0.1,²⁶ a docking program with a physics-based scoring function that performs well relative to the scoring functions of other similar programs (*e.g.* DOCK, FleX, and GOLD⁹⁴⁻⁹⁷). Docking parameters were optimized for the positive-control docking of oseltamivir into the group-1 neuraminidase (N1) crystal structure. The initial AutoDock population size was set to 200 individuals, the maximum number of energy evaluations to 7×10^6 , the number of runs to 25, and the RMSD tolerance to 2.0. All other AutoDock parameters were set to the default values. The AutoDock-predicted binding energy was taken to be the energy associated with the most populated AutoDock cluster. AutoDock grids were calculated for regularly

spaced points at intervals of 0.375 Å contained within a cube 24.00 Å x 27.00 Å x 24.75 Å, centered on the neuraminidase active site.

A Novel Fragment Library Derived from FDA-Approved Compounds

To generate novel compounds, AutoGrow drew upon a new fragment library containing 37 637 redundant fragments derived from FDA-approved compounds using an in house script called *Fragmentizer*. To create this novel fragment library, we first obtained the names of hundreds of FDA-approved compounds by searching Drugs@FDA, supplemented with a list provided by the laboratory of Maurizio Pellecchia. The PDB structures of these compounds were downloaded from drugbank.ca¹¹² and filtered to remove those with molecular weight greater than 700 g/mol. After additional processing, 1 174 drugs remained.

For each compound, *Fragmentizer* first identified all single bonds that could be broken without altering the electronic or geometric configuration of neighboring atoms. The program next generated a second list of all possible bond combinations. Each compound was then decomposed by simultaneously cutting all the bonds of each combination and adding hydrogen atoms to the resulting fragments as needed. Following compound decomposition, all fragments with mass greater than 150 daltons were removed, leaving 37 637 fragments. Redundant fragments were not eliminated. Both *Fragmentizer* and the novel fragment library derived from FDA-approved compounds can be downloaded from http://www.nbcr.net/software/downloads/virtual_lib/.

Post-Processing of AutoGrow-Generated Compounds

As a beta version of AutoGrow was used to generate the ligands, the compounds had to be further processed to correct occasional structural errors. The top ten ligands from each of the three AutoGrow runs were visually inspected. Where the atoms of two distinct fragments were very close, those fragments were bound together to form rings. Where two fragments were mistakenly added via the same scaffold linker hydrogen, extra atoms were removed as needed. Additionally, some sulfur atoms were bound to too many hydrogen atoms. These were eliminated or replaced with oxygen atoms as necessary. Following corrections, each ligand underwent five hundred steps of Cartesian minimization in ICM (Molsoft), a molecular modeling and docking program, prior to being evaluated for drug likeness (Table 3).

Table 3: The initial AutoGrow-generated compounds prior to modification, ranked according to the ensemble-average AutoDock score. The six positive controls are also included.

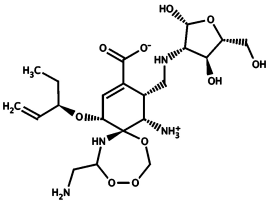
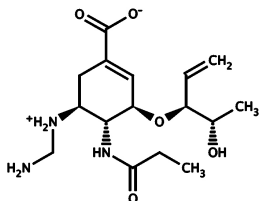
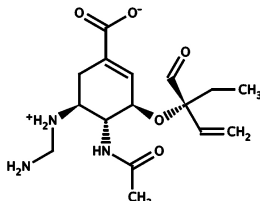
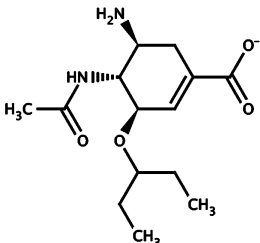
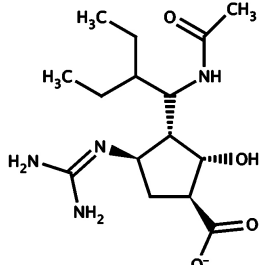
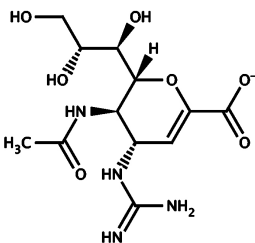
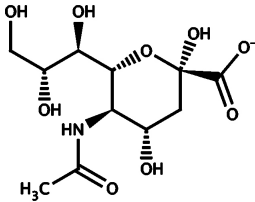
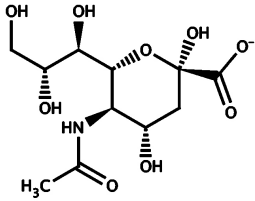
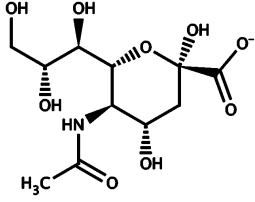
		Ensemble-Average AutoDock Score	Weight	HBA	HBD	LogP
1		-12.83	503.24	12	10	-2.89
2		-11.13	337.19	6	5	-1.71
3		-11.06	335.18	6	4	-1.75
		-8.24	284.17	4	4	2.03
	(Oseltamivir)					
		-7.67				
	(Peramivir)					

Table 3 continued:

	Ensemble-Average AutoDock Score	Weight	HBA	HBD	LogP
 (Zanamivir)	-7.39				
 (SIA Boat)	-6.20				
 (SIA Twist)	-5.89				
 (SIA Chair)	-4.94				

Relaxed Complex Docking

The relaxed complex scheme²⁸ was used to rescore predicted inhibitors. All compounds were docked into the 27 ensemble configurations using AutoDock 4.0.1.²⁶ Additionally, six positive controls were included: sialic acid, the natural

neuraminidase substrate, in the boat, chair, and twist conformations; zanamivir and oseltamivir, FDA-approved neuraminidase inhibitors; and peramivir, a compound currently in clinical trials. An ensemble-average AutoDock score was calculated for each ligand by averaging the AutoDock scores and weighting according to the cluster population size (Table 3):

$$\bar{E} = \frac{\sum_{i=1}^{23} w_i E_i}{\sum_{i=1}^{23} w_i} \quad (1)$$

where $\bar{E}$ is the weighted ensemble-average score, w_i is the size of cluster i , and E_i is the AutoDock score of the ligand docked into the centroid of cluster i .

The docking parameters were unchanged, except that the maximum number of energy evaluations was decreased to 5×10^6 , and the number of runs was increased to 100.

Compound Modification

The top predicted ligand from each of the three AutoGrow runs (compounds 1, 2, and 3; Table 3), as judged by the ensemble-average score, was selected for further examination. Each of these three ligands was loaded into ICM (Molsoft), together with the corresponding ensemble member that gave the best score, and redocked using the ICM docking program (Molsoft). This initial docking established a baseline score that was subsequently used to judge whether modifications to the compound improved or reduced binding affinity.

Each of the three ligands then underwent a series of manual modifications, producing fourteen novel drug-like compounds. Modifications were made such that the ICM docking score was maintained, drug-like metrics (molecular weight, number of hydrogen bond acceptors and donors, predicted LogP, etc.) were improved, chirality was reduced, and molecular rigidity was increased. Furthermore, visual inspection led to the addition of some novel groups that AutoGrow had not suggested. These “modified compounds” were rescored using the relaxed complex scheme, with the same AutoDock and AutoGrid parameters used previously.

4.3 Results and Discussion

The influenza virus has caused several pandemics in the last century; recently, a new pandemic strain (H1N1) has been identified.⁹⁹ In preparation for pandemic influenza, many nations have stockpiled inhibitors of the influenza protein neuraminidase.¹⁰¹ Unfortunately, drug-resistant strains have emerged,¹⁰⁴⁻¹⁰⁷ and the need for novel neuraminidase inhibitors is great.

Neuraminidase has highly flexible loops adjacent to its sialic-acid binding site. In the first crystal structures of group-1 neuraminidase (N1), the so-called 150-loop adopted an open conformation not seen in previous crystal structures of group-2 proteins (N2).¹¹¹ Under different crystallographic conditions, however, the N1 150-loop adopted a closed conformation similar to that of N2.^{113, 114} The flexibility of this loop in N1, as well as implications for drug design, have been further characterized via molecular-dynamics simulations.^{109, 115}

The great flexibility of the N1 active site defies traditional computer-aided drug-design efforts, which typically focus on static crystal structures and at best account for only limited protein flexibility. To aid future drug-design efforts, Cheng et al. recently performed a 40-ns molecular dynamics simulation of the N1 holo enzyme. 1.6×10^4 protein conformations were extracted at regular intervals and clustered into 27 groups using root-mean-square-deviation (RMSD) conformational clustering. The set of the corresponding 27 distinct centroids, representative of all conformations sampled, is said to constitute an *ensemble*.

Initial Compound Generation

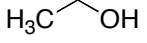
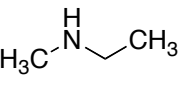
In some fragment-based drug-design strategies, weakly binding molecular fragments, identified experimentally via x-ray crystallography or NMR, or computationally via computer docking, are linked to generate potent composite inhibitors.¹¹⁶ An alternate computational fragment-based growing strategy, exemplified by programs like LUDI¹¹⁷⁻¹¹⁹ and AutoGrow,¹⁰⁸ adds small molecular fragments to initial scaffolds known to bind the target protein, with the goal of improving binding affinity. Because fragment-based strategies are combinatorial, a far greater diversity of compounds can be synthesized and tested than would be possible with traditional high-throughput assays.

To generate predicted ligands, we used a beta version of the computer program AutoGrow,¹⁰⁸ an evolutionary algorithm that automates fragment addition to core scaffolds. Each member of a population of AutoGrow-generated compounds was

evaluated for binding using AutoDock; the best predicted binders became the founding members of the next generation, wherein fragments were again added/modified. Generation after generation, ligands eventually evolved that were well suited to the template protein structures specified. AutoGrow was initially run against an N1 crystal structure (PDB ID: 2HU4¹¹¹). To account for neuraminidase flexibility, the program was run twice more against the top two representative ensemble conformations.

AutoGrow drew upon a fragment library derived from FDA-approved compounds. The library was generated by decomposing approved drugs into their constituent fragments using an in-house script. As AutoGrow selects fragments from the library at random, redundant fragments were maintained to bias the program towards those fragments most commonly found in FDA-approved drugs. For example, the most common fragment (methane) constituted 11.71% of the library. The ten most abundant fragments of this new library are listed in Table 4. Thirty compounds, ten from each AutoDock run, were considered candidates for further analysis.

Table 4: The most common fragments from the library derived by decomposing FDA-approved drugs.

Rank	Compound	Percent of Library
1	CH ₄	11.71%
2	H ₃ C-CH ₃	4.79%
3	H ₂ O	4.28%
4		2.49%
5	H ₃ C-OH	1.93%
6	H ₃ C-NH ₂	1.64%
7	H ₂ C=O	1.40%
8		1.31%
9		1.29%
10		1.28%

Relaxed Complex Rescoring

To further account for N1 flexibility, the compounds were reranked using the relaxed complex scheme.²⁸ In a traditional virtual screen, a docking program is used to predict the ligand energy of binding to a static structure, usually obtained from x-ray crystallography or NMR. The relaxed complex scheme builds upon this traditional methodology by docking candidate ligands into multiple protein conformations extracted from a molecular dynamics simulation. Compounds are then ranked by the

ensemble-average predicted binding energy, rather than by the score associated with a single static structure alone. The relaxed complex scheme has been used to identify inhibitors of FKBP,²⁹ HIV integrase,³⁰ and *T. brucei* RNA editing ligase 1.³¹

In addition to rescoring the thirty AutoGrow-generated compounds, we also scored six positive controls: sialic acid, the natural neuraminidase substrate, in the boat, chair, and twist conformations; zanamivir and oseltamivir, FDA-approved neuraminidase inhibitors; and peramivir, a compound currently in clinical trials (Table 3). The measured IC₅₀ values of oseltamivir, peramivir, and zanamivir are 0.33 ± 0.27 , 0.37 ± 0.26 , and 0.57 ± 0.46 nM, respectively.¹²⁰ These IC₅₀ values are roughly three orders of magnitude smaller than the ensemble-average AutoDock-predicted inhibition constants of 2.25, 5.01, and 8.58 μ M, respectively. However, there is a clear correlation between prediction and measurement ($R^2 = 0.92$), and, importantly, the relaxed complex scheme ranked the known neuraminidase inhibitors correctly.

The experimentally measured K_i of sialic acid suggests that binding to the natural substrate is much weaker. The measured K_i of 50 μ M¹²¹ corresponds well with the ensemble-average AutoDock-predicted inhibition constants of 33.67, 68.02, and 375.48 μ M for sialic acid in the boat, twist, and chair conformation, respectively. When the lowest predicted sialic-acid inhibition constant is considered, the correlation between prediction and measurement is further strengthened ($R^2 = 0.97$).

This strong correlation demonstrates that, while the ensemble-average predicted binding energies and inhibition constants are not numerically accurate, the relaxed complex scheme can be used to accurately rank candidate neuraminidase

inhibitors. Notably, all thirty AutoGrow-generated compounds ranked better than currently approved neuraminidase inhibitors when the relaxed complex scheme was used; the score of the best novel compound was -12.83 kcal/mol, compared to -8.24 kcal/mol for the best FDA-approved inhibitor, oseltamivir.

Compound Modification

Despite the use of fragments derived from FDA-approved drugs, the AutoGrow-generated compounds were not particularly drug like. Consequently, the top ligand from each of the three AutoGrow runs (compounds 1, 2, and 3; Table 3), as judged by the ensemble-average score, was selected for further examination. Fourteen drug-like compounds were subsequently generated by making manual modifications to the three lead ligands. Manual modifications were guided by six rules: satisfy Lipinski's rule of five,⁷⁵ conserve the predicted binding affinity, remove buried unpaired hydrogen bond donors and acceptors, increase molecular rigidity, and reduce chirality.

Lipinski's rule of five. The AutoDock-generated compounds were generally too large, leading to multiple Lipinski violations.⁷⁵ Lipinski's rule of five, developed in 2001, states that compounds generally have poor absorption or permeation if they possess two or more of the following properties: more than five hydrogen-bond donors, more than ten hydrogen-bond acceptors, a molecular weight greater than 500 daltons, or a CLogP greater than five. As the AutoGrow-generated compounds were generally large, most contained too many hydrogen bond donors and acceptors and too

large a molecular weight. Manual modifications reduced the size of the candidate ligands so they would better conform to Lipinski's rule of five.

Predicted binding energy. The reduction in size was generally accompanied by a mild drop in the predicted binding affinity. Some interacting groups had to be removed in order to reduce the molecular weight. Additionally, most steps (i.e. chemical modifications) away from the AutoGrow-generated ligand, which had been "optimized" for binding energy, naturally reduced binding affinity. Nevertheless, the compounds were redocked with each manual modification, and modifications that caused precipitous drops in the predicted binding energy were rejected.

Hydrogen bonds. Some of the docked AutoGrow-generated compounds had hydrogen-bond donors and acceptors that were buried but unpaired (i.e. "unsatisfied"). The hydrogen-bond donors and acceptors of a fully hydrated ligand are typically satisfied via interactions with the water solvent. A large energy penalty occurs upon binding if unsatisfied hydrogen-bond donors and acceptors are positioned in solvent-inaccessible regions along the protein-ligand interface because a hydrogen bond with the water solvent is lost without the compensatory creation of a protein-ligand hydrogen bond. Some of the manual modifications made served to remove these unpaired hydrogen-bond donors and acceptors.

Compound rigidity. Additional manual modifications were made to increase compound rigidity. In the unbound state, non-rigid, "floppy" compounds undergo many different conformational transitions as they twist about their rotatable bonds. Upon binding, however, the ligand is stabilized, resulting in a large entropic penalty of

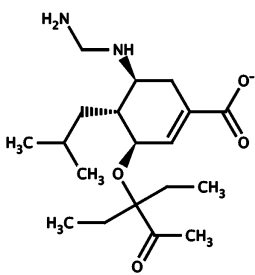
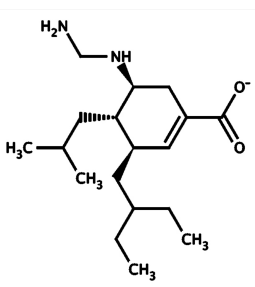
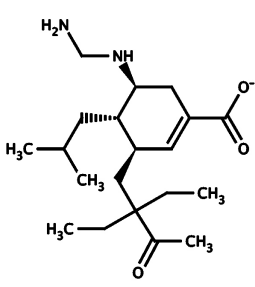
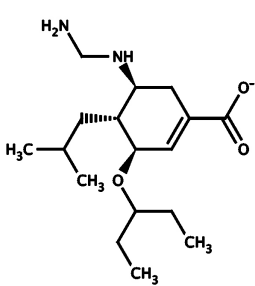
binding. For rigid compounds, the difference in entropy between the bound and unbound state is not as great, and so the penalty is less.

These modified compounds were rescored using the relaxed complex scheme (Table 5). Ten of the fourteen modified compounds ranked better than FDA-approved neuraminidase inhibitors; the ensemble-average score of the best modified compound was -11.05 kcal/mol, compared to -8.24 kcal/mol for oseltamivir.

Analysis of the Top Four Modified Compounds

The top four modified compounds are all similar, in part because they are built on the same cyclohex-1-enecarboxylate scaffold. While similar to oseltamivir, the novel compounds are decorated with different molecular fragments that, in theory, serve to enhance the binding affinity.

Table 5: The top four modified compounds, as judged by the ensemble-average AutoDock score.

		Ensemble-Average AutoDock Score	Weight	HBA	HBD	LogP
4		-11.05	353.24	6	3	2.72
5		-10.96	309.25	4	3	4.65
6		-10.72	351.26	5	3	4.02
7		-10.68	311.23	5	3	3.09

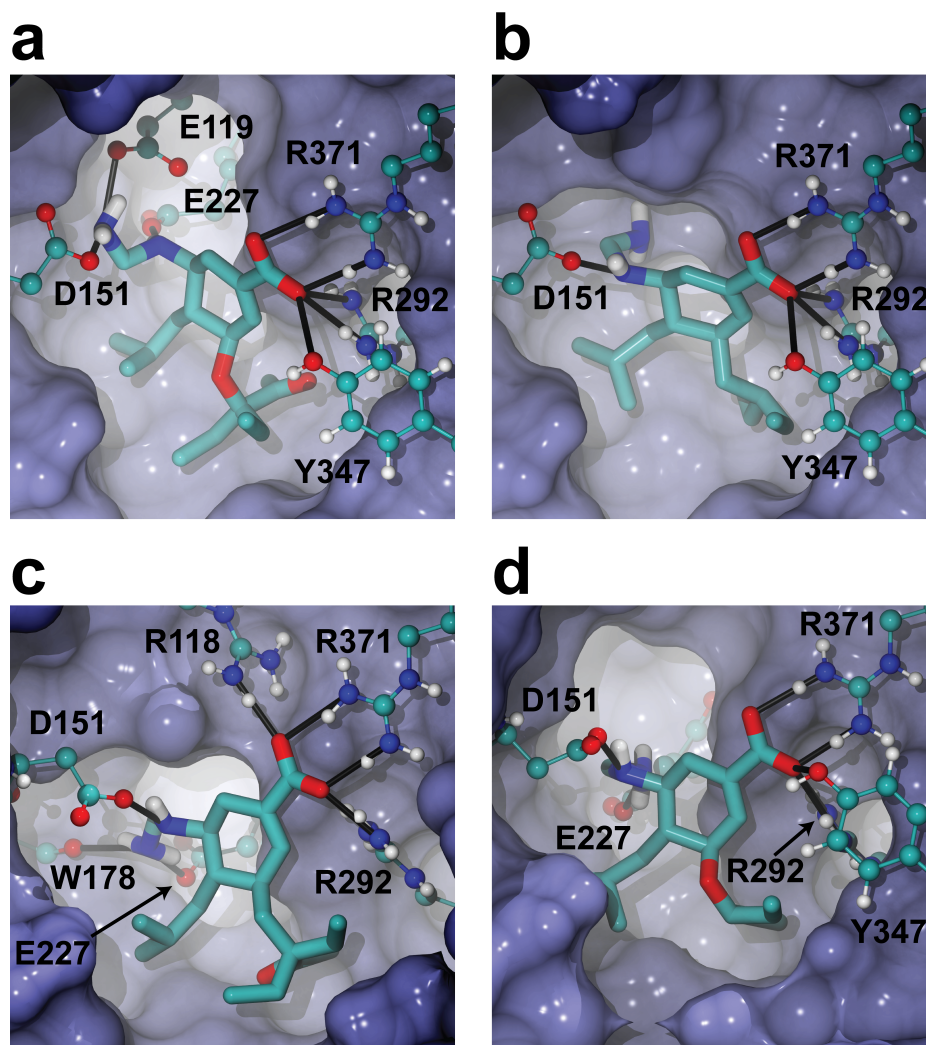


Figure 7: The top four modified compounds docked into the ensemble conformation that gave the best predicted binding energy. Hydrogen bonds are shown as black lines. Some portions of the protein have been removed to facilitate visualization. A) Compound 4. B) Compound 5. C) Compound 6. D) Compound 7.

All four compounds have a carboxylate group in the one position, similar to the known neuraminidase inhibitors oseltamivir, paramivir, and zanamivir. An examination of an oseltamivir-neuraminidase crystal structure¹¹¹ reveals that this carboxylate group participates in hydrogen bonds with R371 and Y347 (distance cutoff, 3.0; angle cutoff, 30°). In addition, an analysis of the top four modified

compounds docked into the 27 ensemble conformations suggests that the carboxylate group also forms hydrogen bonds with R292 and, on rare occasions, with R118 (Figure 7).

All four modified compounds have an (aminomethyl)amino (aminal) group at the five position instead of an amino group, as in oseltamivir. The crystal structure demonstrates that the oseltamivir amine forms hydrogen bonds with D151 and E119. In contrast, an analysis of the top four modified compounds suggests that binding between the aminal fragment and residues D151 and E119 occurs only occasionally. Rather, the distal amine binds instead with E227 and, less frequently, with E277 and the backbone carbonyl of W178. In the case of compounds 5 and 7, the proximal amine is predicted to bind occasionally with E227 as well. Crystal structures indicate that the analogous guanidine groups of peramivir¹²² and zanamivir¹²³ likewise form hydrogen bonds with the backbone carbonyl of W178 (Figure 7).

Unfortunately, aminal groups are unstable and subject to hydrolysis. Initial efforts to find alternative groups failed; when the aminal was changed to a urea, the ensemble-average predicted binding energy was significantly reduced (data not shown). Another solution may be to replace the aminal with a guanidino group, which is essentially a rigid aminal with an additional amine bound to the bridging carbon atom. We again note that zanamivir and peramivir, neuraminidase inhibitors with IC_{50} 's comparable to that of oseltamivir,¹²⁰ have guanidino groups at the analogous location. Alternatively, the proximal aminal nitrogen atom could be replaced with a carbon atom, though such a change may make synthesis more challenging.

The top four modified compounds contain a 2-methylpropyl group at the four position instead of the acetamido group characteristic of oseltamivir, peramivir, and zanamivir. The crystal structures of oseltamivir, peramivir, and zanamivir bound to neuraminidase^{111, 123} suggest that a hydrogen bond may form between the acetamido carbonyl oxygen atom and R152, though in the case of oseltamivir the hydrogen bond is strained (D-H-A angle of 54°).¹¹¹ Additionally, the acetamido methyl group may contribute to the overall binding affinity via hydrophobic interactions with W178. We note, however, that the acetamido amine hydrogen, a hydrogen-bond donor, is buried and unpaired. Like the acetamido group, the 2-methylpropyl group of the top four modified compounds is likewise predicted to form hydrophobic interactions with W178, but without the need for a buried but unpaired hydrogen-bond donor (Figure 7). We note, however, that, in addition to facilitating a hydrogen bond between the carbonyl oxygen atom and the receptor, the amide linker may also simplify chemical synthesis, and so may be necessary on those grounds.

Compounds 4, 5, and 6 have (3-ethyl-2-oxopentan-3-yl)oxy; 2-ethylbutyl; and 2,2-diethyl-3-oxobutyl groups, respectively, at the three position, instead of the pentan-3-yloxy group of oseltamivir. An analysis of the binding poses of the top four modified compounds suggests that the bridging oxygen atom of the oseltamivir pentan-3-yloxy fragment does not contribute significantly to the energy of binding. The crystal structure of oseltamivir bound to neuraminidase reveals no hydrogen bonds with this oxygen atom¹¹¹; additionally, an ensemble-wide analysis of

compounds 4 and 7, which include the bridging oxygen atom, likewise revealed no hydrogen-bond formation (Figure 7).

The importance of the bridging oxygen atom can be further assessed by comparing the predicted binding energies of compounds that differ only at this location. For example, the ensemble-average predicted binding energies of compounds 5 and 7 (Table 5), which differ only in the presence or absence of the bridging oxygen atom, are nearly identical. A comparison between compounds 4 and 6 gives a similar result (Table 5). Though the bridging oxygen is likely energetically unnecessary, we again note that it may greatly facilitate chemical synthesis.

The top four compounds all mimic oseltamivir at the three position in that they contain multiple aliphatic chains. However, compounds 4 and 6 also contain a carbonyl group, a potential hydrogen-bond acceptor. An analysis of these compounds docked into the 27 ensemble conformations revealed that the carbonyl oxygen atom is predicted to form occasional hydrogen bonds with R152. Additionally, the carbonyl oxygen atom of compound 4 forms occasional hydrogen bonds with R292, and the carbonyl oxygen atom of compound 6 forms occasional hydrogen bonds with N294 (Figure 7).

4.4 Conclusion

Pandemic influenza (H1N1) poses an imminent threat, and drug-resistant strains have already emerged.¹⁰⁴⁻¹⁰⁷ In the current work, the computer program AutoGrow¹⁰⁸ was used to generate novel predicted neuraminidase inhibitors. Given

the great flexibility of the N1 active site, protein dynamics were also incorporated into the computer-aided drug-design process. Several potential inhibitors were identified that are predicted to bind neuraminidase better than current FDA-approved drugs.

The neuraminidase inhibitors oseltamivir and zanamivir, both FDA approved, have similar binding poses.^{111, 123} Despite these similarities, these two compounds have very different resistance profiles. As of 2005, no virus resistant to zanamivir had been isolated from an immunocompetent patient, but resistance to oseltamivir was then emergent.¹²⁴ The principal difference between oseltamivir and zanamivir, and the difference likely responsible for their disparate resistance profiles, is the set of molecular fragments used to decorate the central six-member ring. We are therefore hopeful that the novel neuraminidase inhibitors suggested here, with their unique decorating fragments, may likewise serve as scaffolds for future neuraminidase inhibitors against which resistance has not yet developed.

4.5 Acknowledgements

JDD is funded by a Pharmacology Training Grant through the UCSD School of Medicine. This work was carried out with funding from NIH GM31749, NSF MCB-0506593, and MCA93S013 to JAM. Additional support from the Howard Hughes Medical Institute, the National Center for Supercomputing Applications, the San Diego Supercomputer Center, the W.M. Keck Foundation, Accelrys, Inc., the National Biomedical Computational Resource, and the Center for Theoretical Biological Physics is gratefully acknowledged. We would also like to thank Dr. Rommie E. Amaro for helpful discussions and support.

Chapter 4, in full, is a near reprint of the material as it appears in Computational Biology and Chemistry, volume 34, pages 97-105. The dissertation author was the primary co-contributor to this work. J. Andrew McCammon served as the head adviser and principal investigator. Elsevier holds the copyright for this manuscript and has authorized its publication in this dissertation.

CHAPTER 5: Computational Identification of Uncharacterized Cruzain Binding Sites

Jacob D. Durrant, Henrik Keränen, Benjamin A. Wilson, and J. Andrew McCammon

5.1 Introduction

American trypanosomiasis, also known as Chagas disease, is endemic to Central and South America, where 90 to 100 million people are at risk of infection,¹²⁵ 10 to 20 million people are infected,^{14, 125} and 50,000 die annually.¹¹ The disease is caused by the unicellular parasite *Trypanosoma cruzi* (*T. cruzi*), an organism transmitted by insects of the *Reduviidae* family. After drawing a blood meal from its human host, the insect reflexively releases feces containing the parasite into the resulting wound.¹²⁶ Once blood borne, the parasites infiltrate host cells and replicate. Following replication and maturation, host cells burst open, releasing new *T. cruzi* parasites into the bloodstream.¹²⁷

The acute phase of the disease, which typically persists for two months and has a fatality rate of 2 to 8%, is characterized by the mononuclear inflammation and necrosis of parasitized cells, especially in the heart.¹²⁸ The chronic stage of the disease is characterized by dilated cardiomyopathy; indeed, American trypanosomiasis is the leading cause of infectious myocarditis in the world.¹²⁹

New therapies for Chagas disease are urgently needed. Current treatments, nitrofurans like nifurtimox and benznidazole, are highly toxic,^{16, 128, 130} and drug resistance has been reported.¹³¹ Furthermore, one recent study demonstrated that these compounds neither eradicate the parasite nor prevent cardiomyopathy over the long term.¹³²

The major cysteine protease of *T. cruzi*, called cruzain or, alternatively, cruzipain, is one attractive drug target.¹² A member of the peptidase C1 protein family, cruzain is present and essential in all stages of *T. cruzi* development.^{13, 14} Over-expression of cruzain enhances the transformation of the parasite into the infective form,¹⁵ and reduced protease activity prevents infection in wild-type mice.¹⁶ Additionally, cysteine protease inhibitors block both the replication and the differentiation of the parasite *in vitro* and *in vivo*.^{12, 17-24} Cruzain inhibitors can cure infection in cell, mouse, and dog models.^{20, 25}

The future rational design of improved cruzain inhibitors necessitates a better understanding of the flexibility and conformational changes characteristic of the cruzain active site. Molecular dynamics (MD) simulations, in which the forces that act on the atoms of a molecular system are approximated using Newton's laws of motion, can be powerful tools for better understanding protein flexibility and conformational sampling relevant to drug design. For example, one recent MD study of HIV integrase revealed a previously uncharacterized binding trench that was subsequently exploited in the design of Isentress (raltegravir), an HIV drug approved by the FDA in 2007.³⁰ Importantly, this trench was not evident in the then available crystal structures; it was only by studying active-site flexibility *via* MD that the trench was initially identified.

Additional novel sites of enzymatic, allosteric, or structural importance can be identified computationally by comparing the sequence of the target protein with evolutionarily related enzymes. Critical protein residues are often conserved across multiple members of the same protein family; once multiple sequences are aligned,

conserved patches of protein residues can be easily identified. Additional experimental studies can then characterize the pharmacological significance of these patches.

Given the urgent need for novel antichagastic therapeutics, we here use computational methods, including molecular dynamics (MD) simulations and a sequence alignment of a non-redundant, unbiased set of peptidase C1 family members, to identify previously uncharacterized binding regions that may serve as sites for future pharmacological intervention.

5.2 Methods

MD Simulations

To prepare cruzain for MD simulations, hydrogen atoms were added to a high-resolution cruzain crystal structure (PDB: 1ME4¹³³) using PDB2PQR to approximate protein protonation at pH 5.5, the pH of the reservosome where cruzain is located in the epimastigote stage of the parasite.^{49, 134, 135} Protonation states were subsequently verified manually. Hydrogen atoms were added to the bound hydroxymethyl-ketone inhibitor using Discovery Studio (Accelrys). The LEaP module of the AMBER9 suite⁵⁴ was used to solvate the system by submerging the protein in a TIP3P water box⁵⁵ that extended 10 Å beyond the protein in all directions. All crystallographic water molecules were maintained. Ten sodium cations were added to make the system electrically neutral; additional ions were then added to simulate a more physiological 20 mM NaCl solution. The system was parameterized using the generalized and FF99SB AMBER force fields.^{136, 137}

NAMD2.7b1⁵⁹ was used for all MD simulations. Periodic boundary conditions were employed with the particle mesh Ewald method to account for electrostatic effects (smoothing cutoff: 14 Å). Langevin dynamics were applied to maintain the temperature, and a modified Langevin piston Nosé-Hoover thermostat was used to maintain 1 atm pressure. The initial structure was minimized in four distinct steps; hydrogen atoms were first relaxed for 5,000 steps; hydrogen atoms, water molecules, and ions were next relaxed for 5,000 steps; hydrogen atoms, water molecules, ions, and protein side chains were then relaxed for 10,000 steps; and, finally, all atoms were relaxed for 25,000 steps. Following minimization, the system was equilibrated with an NPT-ensemble at 310 K using stepwise harmonic-constraint force constants of 4, 3, 2, and 1 kcal/mol/Å² on the protein backbone. 250,000 steps of MD simulation were executed for each force constant (1 fs time step).

Following minimization and equilibration, five distinct 20-ns productive runs were performed (10^7 steps of 2 fs) with distinct random seeds in order to sample many protein configurations.

Trajectory Clustering

The RMSD-based *gromos* clustering algorithm, as implemented in the GROMACS++ computer package (*g_cluster*), was used to cluster the conformations sampled during the five 20-ns MD simulations.⁶¹ Structures were first extracted from the trajectories every 50 fs, generating 4,002 snapshots total. These snapshots were aligned by their C_α atoms and clustered on the 73 residues of the cruzain active site,

defined as all residues within 10 Å of the ligand: 18-31, 50, 53-54, 57-72, 74, 91, 93-98, 115, 117, 120, 136-142, 144-145, 158-165, 181-184, 203-210.

The *gromos* clustering algorithm was first described by Daura et. al.¹¹⁰ In brief, for each protein conformation in a pool of conformations, the RMSD distance between the atoms of the aforementioned residues and the corresponding atoms of every other protein conformation in the pool (potential “neighbors”) is calculated. The conformation with the most neighbors within a user-specified distance cutoff (“close neighbors”) is then selected. This conformation, together with its close neighbors, constitutes the first cluster. The protein conformations of the first cluster are then removed from the pool, and the process is repeated with the remaining conformations until none are left.

When a cutoff of 0.95 Å was used, this procedure produced 24 clusters. The central member of each cluster was considered most representative; the set of all central members is said to constitute an *ensemble*.

Calculating Beta Factors from the MD Simulation

To derive beta factors from the motions sampled during the MD simulations, all trajectories were concatenated, and the RMSF of each protein residue was calculated using the AMBER 9 *ptraj* module.⁵⁴ These RMSF values were converted into beta factors by multiplication, where $\beta = \text{RMSF} * 8\pi^2 / 3$.

Virtual Screening

A small-molecule library was prepared from the ligands of the NCI Diversity Set II, a set of freely available, diverse, drug-like molecules. The Schrödinger LigPrep program (Schrödinger) was used to assign protonation states at pH 5.5 and to identify and generate tautomers and stereoisomers. One ligand could not be processed with LigPrep; instead, hydrogen atoms were added to this ligand and its geometry was optimized using Discovery Studio (Accelrys).

The ligands of this small-molecule library were docked into a 1.20 Å cruzain crystal structure (PDB ID: 1ME4¹³³). Hydrogen atoms were added using PDB2PQR^{49, 135} at pH 5.5. At this pH, C25 and H159 formed the thiolate/imidazolium pair required for the catalytic mechanism.¹³⁸ An initial virtual screen was performed using the CDOCKER docking software (Accelrys) with a docking sphere 15 Å in diameter centered on the coordinates of the crystallographic ligand, as that program was able to recapture the crystallographic poses of two known hydroxymethyl-ketone cruzain inhibitors.¹³³ The CDOCKER-predicted pose of each of the ligands was rescored using the PLP2 scoring function.¹³⁹ The best ligands as evaluated by PLP2 were compiled into a new small-molecule ligand library enriched for potential cruzain inhibitors.

To account for receptor flexibility, we subsequently used the relaxed-complex scheme,²⁸ a protocol that has been used previously to identify inhibitors of FKBP,²⁹ HIV integrase,³⁰ and *T. brucei* RNA editing ligase 1.³¹ The compounds of the enriched small-molecule library were docked into the 24 members of ensemble, again using CDOCKER (Accelrys). Each of these compounds was rescored with the PLP2

¹³⁹ scoring function. For each ligand, a PLP2-based ensemble-average score was calculated according to the following equation:

$$\overline{E} = \frac{\sum_{i=1}^{23} w_i E_i}{\sum_{i=1}^{23} w_i} \quad (1)$$

where $\overline{E}$ is the weighted ensemble-average score, w_i is the size of cluster i , and E_i is the best score of the ligand, independent of tautomeric or stereoisomeric form, docked into the centroid of cluster i .

Alignment of a Non-Redundant, Unbiased Set of Peptidase C1 Family Members

Cruzain was compared to other members of the peptidase C1 family. First, the UniProt database¹⁴⁰ was used to identify reviewed members of the peptidase C1 family, as defined by the MEROPS classification,¹⁴¹ that had structures deposited in the Protein Data Bank.¹⁴² All amenable sequences except those of cruzain were then aligned using ClustalW in the MultiSeq extension of VMD.¹⁴³⁻¹⁴⁵ A non-redundant set was selected from these aligned peptidase C1 sequences (sequence QR: 75; GF: 1.0). Gaps in the sequences were then removed, and ClustalW was used to align the corresponding sequences to a cruzain crystal structure (PDB: 1AIM) that was chosen as a non-redundant structure from the set of all cruzain structures aligned using STAMP.¹⁴⁶ The following sequences were aligned: 1A6R, 1AEC, 1CJL, 1DEU, 1FWO, 1JQP, 1K3B, 1M6D, 1PCI, 1XKG, 2C0Y, 2CB5, 2FO5, 2O6X, 2WBF,

3PBH, 7PCK, and 8PCH. Residues were colored by similarity according to the BLOSSUM30 matrix.

5.3 Results / Discussion

Motivated by the urgent need for novel antichagastic therapeutics, we set out to identify previously uncharacterized cruzain sites that might serve as future targets for pharmacological intervention. Five 20-ns MD simulations were first used to probe the dynamics of the cruzain active site, as knowledge of protein dynamics can provide important structural insights beyond the information that can be obtained from crystal structures alone.

System Equilibration

While four of the five 20-ns MD simulations equilibrated, as judged by convergent RMSD values, the RMSD plot of the first simulation suggested that several conformational states had been sampled (Figure 8). A careful examination of the trajectory revealed that a mobile N-terminal tail was entirely responsible for the non-convergent RMSD values of the first simulation. In the crystal structure (PDB: 1ME4¹³³), as in four of the five MD simulations, the N-terminal tail is held against the protein *via* hydrogen bonds between the A3 backbone carbonyl and the D167 backbone amine, and between the P2 backbone carbonyl and the Y166 side-chain hydroxyl group. In the first MD simulation, however, the hydrogen bond between P2 and Y166 broke after 6.7 ns. After 14.3 ns, the bond between A3 and D167 broke,

allowing the N-terminal tail to rotate such that new hydrogen bonds were formed between the D167 side-chain carboxylate group and the backbone amines of both A3 and A4. After 18.7 ns, the N-terminal tail returned to its original position. While these conformational changes are interesting, they occur far from the peptide binding site and so are probably not relevant to drug design. Importantly, when the first three residues of the protein are omitted from the RMSD calculation, the RMSD plot of the first MD simulation is convergent, similar to the RMSD plots of the other four simulations.

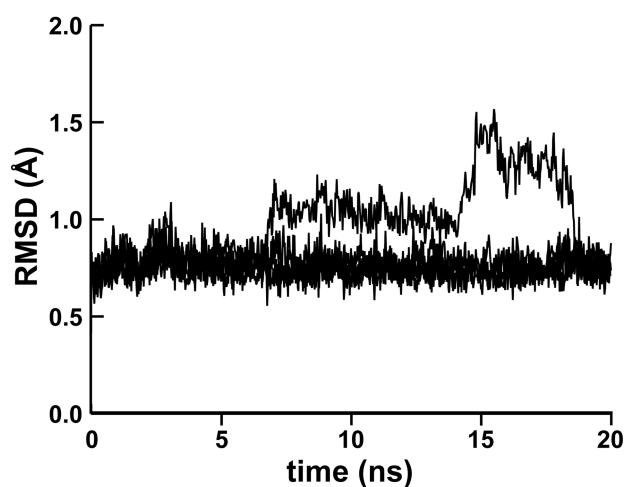


Figure 8: MD equilibration. The five MD trajectories were aligned by protein alpha-carbon atoms, and the RMSD of each trajectory relative to the first frame was calculated as a function of time. The protein N-terminal tail of the first simulation assumed several rotameric states. The remaining four simulations were equilibrated, as demonstrated by their convergent RMSD values.

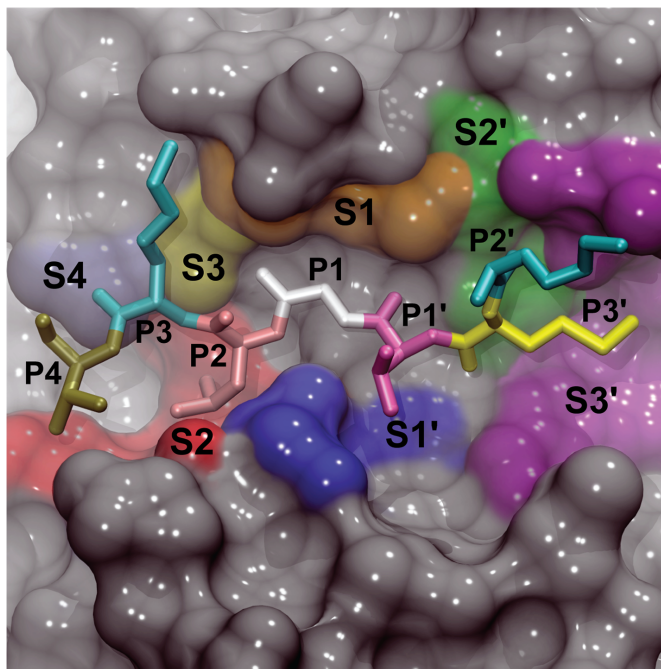


Figure 9: The cruzain active site colored according to the seven substrate-binding subsites. Four subsites on the acyl side of the cleaved peptide bond, named S4, S3, S2, and S1, bind the peptide amino acids P4, P3, P2, and P1. Three subsites on the amino side of the bond, named S1', S2', S3', bind the peptide amino acids P1', P2', and P3'. As no crystal structure of cruzain bound to a peptide substrate was available, the peptide shown was taken from a crystal structure of the homologous protein procathepsin K (PDB: 1BY8).

The Cruzain Active Site

The MD simulations were subsequently used to study the flexibility of the cruzain active site. Cruzain, like other cysteine proteases, contains seven subsites that bind peptide amino acids. Four subsites on the acyl side of the cleaved peptide bond, named S4, S3, S2, and S1, bind the peptide amino acids P4, P3, P2, and P1. Three subsites on the amino side of the bond, named S1', S2', S3', bind the peptide amino

acids P1', P2', and P3' (Figure 9).¹⁴⁷ The only well defined subsites of these seven are S2, S1, and S1', and only S2 and S1' demonstrate significant specificity.¹⁴⁸

To judge the flexibility of the cruzain active site, the beta factor of each protein residue was calculated from the molecular motions sampled during the MD simulation. In general, the active site was remarkable for its great stability, likely in part due to the bound hydroxymethyl-ketone inhibitor.¹³³

Trajectory Clustering

To better distinguish between the many conformational states sampled by the MD simulations, 4,002 protein configurations were extracted from the simulations at regularly spaced intervals and grouped into 24 clusters by RMSD using the *gromos* clustering algorithm.¹¹⁰ The centroid member was selected from each cluster, and the set of all centroid members, representative of the many conformations sampled by the MD simulations, is said to constitute an *ensemble*.

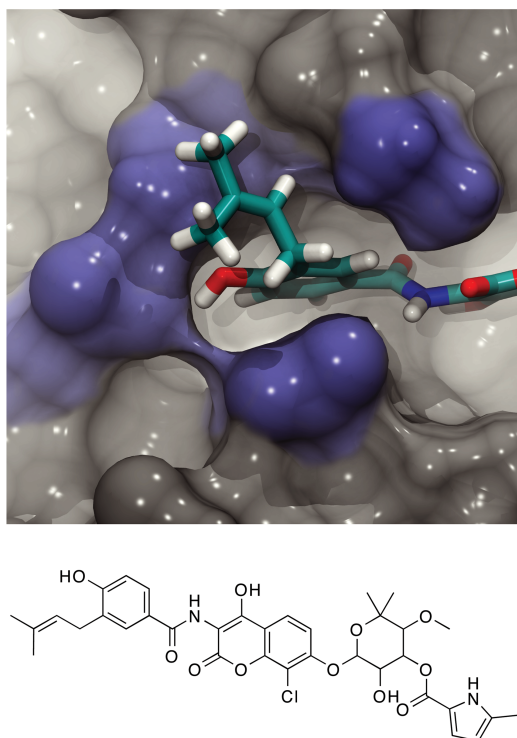


Figure 10: The binding of compound 1 to the semi-open conformation of cruzain. Compound 1 (clorobiocin, bottom), a known antagonist of *T. cruzi* amastigote growth, is predicted to occupy a previously uncharacterized binding pocket adjacent to the S2 subsite (top).

Virtual Screening

To test the potential physiological relevance of the ensemble-member active-site conformations, CDOCKER (Accelrys) was used to dock the compounds of the NCI Diversity set II, a set of freely available, diverse, drug-like molecules, into both the cruzain crystal structure and the 24 protein conformations of the ensemble. A full account of the results of this virtual screen is forthcoming; however, one of the predicted inhibitors warrants further discussion here. Compound **1** (clorobiocin, Figure 10) was the best predicted novel cruzain inhibitor as evaluated by the PLP2 scoring function¹³⁹ in both the screen against the static crystal structure and the relaxed-complex screen against the ensemble of 24 conformations. As positive controls, two

hydroxymethyl-ketone inhibitors (PDB: 1ME3¹³³) were included in the relaxed-complex screen. After rescoring with an ensemble-average PLP2 score, these compounds ranked even better than compound **1**, confirming that the PLP2 scoring function is well suited to this particular protein receptor.

We note with interest that previous studies have demonstrated that compound **1** antagonizes *T. cruzi* amastigote growth.^{71, 149} The primary protein target of clorobiocin is thought to be *T. brucei* topoisomerase II, but the idea of a polypharmacophoric mechanism that includes cruzain inhibition is interesting. The PLP2 scores of compound **1** docked into the central members of the first, second, and third most populated clusters were 95.07, 117.7, and 115.68, respectively.

To understand why compound **1** binding to the second ensemble conformation was favored, the pose of the ligand docked into that conformation was analyzed. While docking poses should never be blindly accepted, this particular pose seemed promising. Aside from having the best PLP2 score, this binding mode placed a conjugated ring in the S2 pocket, similar to the binding modes of some known ligands (e.g. some vinyl sulfone inhibitors⁷⁴) and of some native substrates.¹⁵⁰ Importantly, the docked pose also suggested that one of the ligand rings binds in a previously uncharacterized, druggable pocket immediately beyond the S2 subsite (Figure 10).

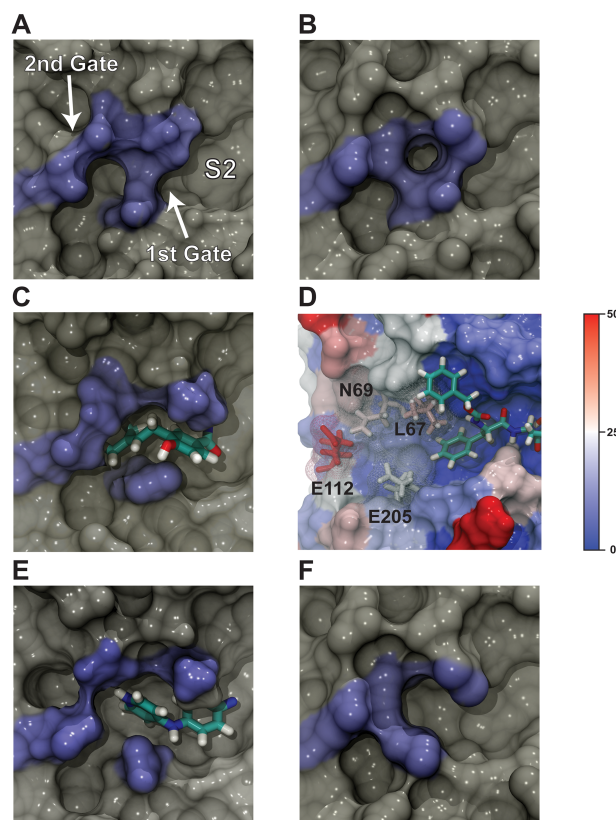


Figure 11: A previously uncharacterized binding pocket adjacent to the S2 subsite. The protein conformations depicted in A, B, C, E, and F are shown from the same orientation. Molecular fragments were excised from selected members of the NCI diversity set docked into semi-open and fully open protein conformations extracted from the MD simulation. A) The central member of the first cluster extracted from the MD simulation, in which the pocket is closed. The S2 subsite, as well as the first and second gate, are labeled. B) A cruzain crystal structure (PDB: 2EFM) in which the pocket is likewise closed. C) The central member of the second cluster, in which the pocket is semi-open. A small molecular fragment is shown docked into the pocket to demonstrate druggability. D) The cruzain active site with residues colored according to beta values calculated from the MD simulation. Blue indicates stability, and red indicates mobility. Several of the residues that comprise the previously uncharacterized pocket are flexible. E) The central member of the third cluster, in which the pocket is fully open. A small molecular fragment is shown docked into the pocket to demonstrate druggability. F) A cruzain crystal structure (PDB: 1ME3) in which the pocket is semi-open.

An Additional Binding Pocket Beyond the S2 Subsite

The beta factors of the protein residues that from this previously uncharacterized pocket revealed significant protein flexibility. Two of the residues that form the distal wall of the S2 subsite, L67 and E205, were somewhat flexible (Figure 11D), and two other protein residues beyond the S2 subsite, N69 and E112, were also mobile (Figure 11D). Together, these four flexible residues comprise two “gates” (L67-E205 and N69-E112) that, when open, form the walls of a previously uncharacterized druggable pocket that medicinal chemists have yet to exploit.

Published cruzain crystal structures hint at the existence of this additional pocket. A crystal structure of cruzain bound to a vinyl sulfone derived inhibitor (PDB: 2EFM) demonstrates a closed configuration (Figure 11B), while a crystal structure of cruzain bound to a hydroxymethyl-ketone inhibitor (PDB: 1ME3¹³³) demonstrates a semi-open configuration (Figure 11F).

The crystal structures, however, do not fully capture the extent of opening demonstrated by the MD simulations. The central member of the top cluster, which accounted for 82.5% of the trajectory, had a closed conformation (Figure 11A). The central member of the second cluster, accounting for 6.9% of the trajectory, had a semi-open conformation (Figure 11C), and the central member of the third cluster, accounting for 3.6% of the trajectory, was fully open (Figure 11E). As shown in Figures 11C and 11E, molecular docking demonstrates that both the semi-open and the fully open conformations can easily accommodate small molecular fragments.

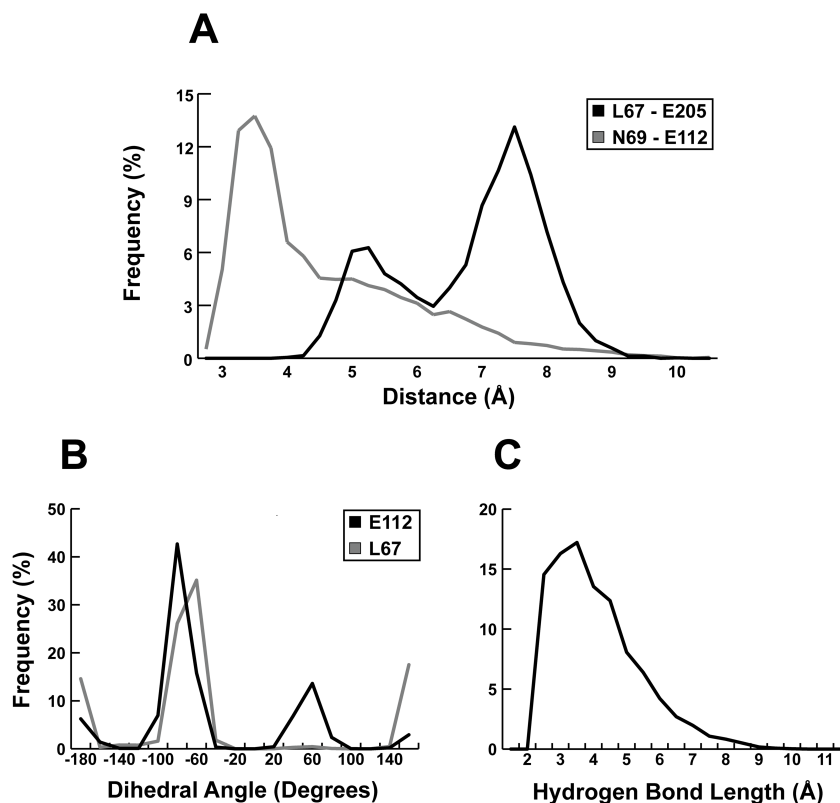


Figure 12: The opening and closing of the two “gates” that comprise the previously uncharacterized pocket. A) Histograms of the distances between the L67 γ carbon atom and the E205 δ carbon atom (the first gate, in black), and the distances between the N69 side-chain amino nitrogen atom and the E112 δ carbon atom (the second gate, in gray), over all 100 ns of trajectory. When the L67-E205 distance is ~ 7.5 Å, the first gate is open. When ~ 5.5 Å, the gate is closed. Similarly, when the N69-E112 distance is ~ 6.0 Å, the second gate is open; when ~ 3.5 Å, the gate is closed. B) Histograms of the dihedral angles defined by the backbone amino nitrogen atom and the α , β , and γ carbon atoms of L67 (in gray) and E112 (in black), respectively, measured over the course of the trajectory. The first gate is open when the L67 dihedral angle is $\sim 60^\circ$ and closed when $\sim 180^\circ$. The second gate is open when the E112 dihedral angle is 60° or 180° , and closed when -60° . C) A histogram of the distances between the N69 side-chain amino nitrogen atom and the E112 carboxylate oxygen atoms, demonstrating the formation of a transient hydrogen bond.

The First Gate

To further characterize the opening and closing of the first gate, the distance between the L67 γ carbon atom and the E205 δ carbon atom (d_I) was monitored over all 100 ns of trajectory. A histogram of these distances was bimodal (Figure 12A) and suggested that the gate was open ($d_I > 6.25$ Å) 70% of the time ($d_I = 7.6$ Å $\pm$ 0.6) and closed ($d_I < 6.25$ Å) 30% of the time ($d_I = 5.4$ Å $\pm$ 0.4). As a reference, this same distance is 7.8 Å and 8.4 Å in the semi-open and fully open conformations, respectively, both of which can accommodate a ligand (Figures 11C and 11E).

We note that the presence of the hydroxymethyl-ketone inhibitor may have affected the dynamics of the first gate by largely immobilizing the E205 residue in an open conformation. E205 plays a unique role in substrate binding. The cruzain S2 subsite, like that of cathepsin B, differs from other cysteine proteases in that it can bind both hydrophobic and basic amino acids.^{14, 150} The E205 residue acts as a highly mobile switch. When a basic amino acid like arginine occupies the S2 subsite, the acidic E205 carboxylate group swings into S2 to interact with the guanidino group, a conformation that can be seen in the crystal structure of cruzain bound to benzoyl-arginine-alanine-methyl ketone (PDB: 2AIM^{14, 150}). When a hydrophobic amino acid occupies the S2 subsite, E205 rotates away from S2 and interacts with the solvent,^{14, 150} a conformation evident in the crystal structure of cruzain bound to WRR-99 (PDB: 1EWL).

The hydrophobic phenyl group of the hydroxymethyl-ketone inhibitor present in the S2 subsite of the MD simulation locked E205 in the open, solvent exposed

conformation (Figure 11D). Consequently, the dynamics of the first gate were mostly determined by L67. A histogram of the dihedral angle defined by the backbone amino nitrogen atom and the α , β , and γ carbon atoms of L67, measured over the course of the trajectory, demonstrated that the side chain of this important residue rotated freely (Figure 12B). Visual inspection confirmed that gate opening occurred when the dihedral angle (θ_i) was roughly -60° ($d_i = 7.6 \text{ \AA} \pm 0.7$ when $-140^\circ < \theta_i < 60^\circ$), and that gate closing occurred when the dihedral angle was roughly 180° ($d_i = 5.7 \text{ \AA} \pm 0.7$ when $\theta_i < -140^\circ$ or $\theta_i > 60^\circ$). By this metric, the first gate was open 67% of the time, a value that matches that found by measuring the distance between the L67 γ carbon atom and the E205 δ carbon atom directly.

The Second Gate

To assess the opening and closing of the second gate, the distance between the N69 side-chain amino nitrogen atom and the E112 δ carbon atom (d_2) was monitored over all 100 ns of trajectory. A histogram of these distances was again bimodal (Figure 12A) and suggested that the second gate is open ($d_2 > 4.5 \text{ \AA}$) 43% of the time ($d_2 = 6.0 \text{ \AA} \pm 1.1$), and closed ($d_2 < 4.5 \text{ \AA}$) 57% of the time ($d_2 = 3.7 \text{ \AA} \pm 0.4$). As a reference, this same distance is $5.3 \text{ \AA}$ in the fully open conformation, which can accommodate a ligand (Figure 11E).

Both N69 and E112, which form the second gate, are mobile. Of these two residues, E112 is particularly flexible (Figure 11D). Visual inspection of the trajectory confirmed that N69 and E112 interact with each other *via* a transient hydrogen bond

between the N69 side-chain amino nitrogen atom and the E112 carboxylate oxygen atoms (Figure 12C). A hydrogen bond between these two residues (distance cutoff of 3.5 Å) was present in roughly 30% of the frames extracted from the trajectory.

As this hydrogen bond was transient, E112 often flipped out into the solvent, where the carboxylate group interacted with water molecules. A histogram of the dihedral angle defined by the backbone amino nitrogen atom and the α , β , and γ carbon atoms of E112 (θ_2), measured over the course of the trajectory, confirmed that the side chain of this important residue can freely rotate (Figure 12B). Visual inspection demonstrated that gate opening occurred at two rotameric states, when the dihedral angle was 60° or 180° ($d_2 = 5.9 \text{ Å} \pm 1.4$ when $\theta_2 > 0^\circ$ or $\theta_2 < -140^\circ$). Gate closing occurred when the dihedral angle was -60° ($d_2 = 4.1 \text{ Å} \pm 0.9$ when $-140^\circ < \theta_2 < 0^\circ$). By this metric, the second gate was open 34% of the time.

Types of Fragments that Bind this Previously Uncharacterized Pocket

To determine what kinds of molecular fragments would best fit into the previously uncharacterized pocket immediately beyond the S2 subsite, we examined the predicted binding poses of NCI compounds docked into the third (fully open) ensemble conformation (Figure 11E). Roughly two-dozen ligands were predicted to occupy the previously uncharacterized pocket and to bind cruzain with high affinity. With some exceptions, the molecular fragments occupying the previously uncharacterized pocket were generally aromatic rings or aliphatic chains, often with hydroxyl groups that formed hydrogen bonds with the E205 carboxylate oxygen

atoms. Numerous FDA-approved drugs contain hydroxylated rings (e.g. masoprocol, carbidopa, acetaminophen, etc.) and/or aliphatic chains (e.g. penciclovir, ethambutol, and miglitol), and so these fragments can be considered drug like.

Alignment of a Non-Redundant, Unbiased set of Peptidase C1 Family Members

Having used MD simulations to identify a previously uncharacterized binding pocket immediately beyond the S2 subsite, we next used a bioinformatics approach to identify other possible sites of importance. Residues critical to protein function, like those of an enzymatic or allosteric active site, like those that participate in essential protein-protein interactions, or like those that play important structural roles, are often conserved across multiple homologous members of the same protein family. To identify these critical residues, cruzain was compared to other members of the peptidase C1 family.

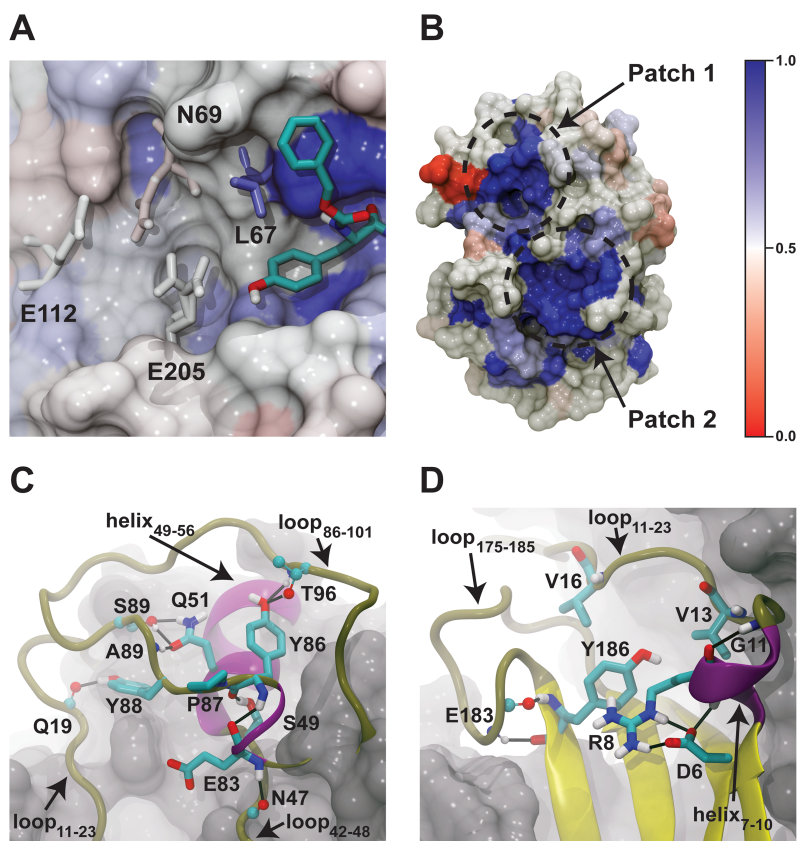


Figure 13: Cruzain (PDB: 1AIM) colored by residue conservation across multiple members of the peptidase C1 family. Conserved residues are shown in blue, and nonconserved residues are shown in red. A) The residues of the seven peptide-binding subsites are generally conserved. The S2 subsite, critical for specificity, is an important exception. Additionally, one of the residues of the previously uncharacterized binding pocket immediately beyond the S2 subsite, L67, is well conserved, while the remaining three residues of the pocket, N69, E112, and E205, are not. B) Two patches of highly conserved residues on the side of the protein opposite the proteolytic active site can be seen. The first, patch one, is comprised of Y88, P87, E83, Y86, Q51, and S49. The second, patch two, is comprised of Y186, R8, V16, G11, D6, and V13. C) A close-up view of the first patch. Conserved residues are shown in licorice, and non-conserved residues are shown in balls and sticks. Side chains or backbone atoms that do not participate in hydrogen bonds have been removed for the sake of clarity. D) A similar close-up view of the second conserved patch.

As expected, the residues of the seven subsites of the proteolytic binding pocket are generally conserved (Figure 13A). The S2 subsite, critical for specificity, is an important exception; this site, like the S2 subsite of cathepsin B, differs from other cysteine proteases in that it can bind both hydrophobic and basic amino acids.^{14, 150}

Additionally, the six cysteine residues involved in disulfide bonds are conserved, suggesting that these bonds are critical for protein tertiary structure. A natural mutation in human cathepsin C, a related cysteine protease with the same papain fold, confirms this importance. Patients with a cathepsin C C291Y mutation, equivalent to a cruzain C56Y mutation, develop Papillon-Lefèvre syndrome due to cathepsin C dysfunction.¹⁵¹

Surprisingly, there are two patches of highly conserved residues on the side of the protein opposite the proteolytic active site. The first, patch one, is comprised of Y88, P87, E83, Y86, Q51, and S49. The second, patch two, is comprised of Y186, R8, V16, G11, D6, and V13 (Figure 13B). Both of these patches lie in a long shallow groove, formed largely by several disordered loops, which traverses the protein surface (Figure 13B). These loops include the loops spanning G11 to G23 (loop₁₁₋₂₃), G42 to S48 (loop₄₂₋₄₈), Y86 to T101 (loop₈₆₋₁₀₁), and N175 to G185 (loop₁₇₅₋₁₈₅). Though disordered, these loops are held rigid by the conserved residues of the two patches, which bind the loops to stable tertiary structures and/or to each other. This rigidity may serve to maintain the shape of the traversing groove.

To the best of our knowledge, this groove and its associated conserved patches, which are common to members of the peptidase C1 family, have not been previously

characterized. These highly conserved patches may play roles in allosteric regulation or structural stability. Additionally, the shallow traversing groove formed by these two patches may also constitute a surface amenable to protein binding.

We first turned our attention to the first patch of conserved residues. S49 and Q51 are highly conserved buried residues that belong to a stable helix spanning S49 to L56 (helix₄₉₋₅₆). Interactions between these residues and residues of loop₈₆₋₁₀₁ help to pin the loop against the stable helix, thereby imparting stability to part of the traversing groove. The S49 side-chain hydroxyl group hydrogen bonds with the backbone carbonyl oxygen of Y86, another conserved residue. Additionally, the side-chain carbonyl oxygen atom of Q51 forms two hydrogen bonds, one with the backbone amine of A89 and one with the side-chain hydroxyl group of S89. The side-chain amine of Q51 hydrogen also hydrogen bonds with the side-chain hydroxyl group of S89 (Figure 13C).

Though cruzain mutagenesis data is absent from the literature, studies of the closely related human cathepsin C protein likewise suggest that S49 and Q51 have important roles. Patients with a cathepsin C S284N mutation, analogous to a cruzain S49N mutation, develop Papillon-Lefèvre syndrome,¹⁵² and patients with a cathepsin C Q286R mutation, analogous to a cruzain Q51R mutation, develop Haim-Munk syndrome. Both these syndromes are caused by cathepsin C dysfunction.¹⁵³

Y86 and P87, also conserved residues of the first patch, likewise seem to play an important role in imparting rigidity to the disordered loop₈₆₋₁₀₁. The Y86 side-chain hydroxyl group forms two hydrogen bonds with T96, helping to maintain the hairpin

shape of loop₈₆₋₁₀₁. P87 does not participate in any hydrogen bonds, but the conformational rigidity of the proline backbone may contribute to the overall rigidity of loop₈₆₋₁₀₁ as well (Figure 13C).

The rigidity of loop₈₆₋₁₀₁ is in part transferred to loop₁₁₋₂₃ and loop₄₂₋₄₈ *via* the conserved residues Y88 and E83, respectively. The Y88 side-chain hydroxyl group hydrogen bonds with the Q19 backbone carbonyl oxygen atom, and the E83 backbone amine hydrogen bonds with the N47 backbone carbonyl oxygen atom, thereby holding all these loops rigid relative to one another. It is also interesting to note that the side-chain carboxylate group of N47 is solvent exposed and potentially capable of interacting with other proteins or small-molecule compounds that may bind in the traversing groove (Figure 13C).

We next turned our attention to the second patch of conserved residues. The conserved residues of this patch likewise serve to hold disordered loops rigid against underlying secondary structures. For example, a hydrogen bond exists between the backbones of two highly conserved residues, G11 and R8, that anchors part of loop₁₁₋₂₃ to a small helix spanning W7 to R10 (helix₇₋₁₀). Helix₇₋₁₀ is in turn positioned relative to an underlying beta sheet by multiple hydrogen-bond interactions between the conserved residues R8 and D6. These interactions are likely critical for protein function; D6 is analogous to the cathepsin C residue D236, and patients with D236Y mutations develop Papillon-Lefèvre syndrome, again suggesting cathepsin C dysfunction (Figure 13D).¹⁵¹

The conserved residues of the second patch also impart some structure and rigidity to loop₁₇₅₋₁₈₅. The backbone of the conserved residue Y186, part of a stable underlying beta sheet, forms two hydrogen bonds with the backbone of E183. These interactions not only hold part of loop₁₇₅₋₁₈₅ fixed relative to the beta sheet, but also help stabilize a sharp turn at the sheet-loop junction. Additionally, the phenol group of Y186 forms an interesting π -cation interaction with R8, also conserved. This interaction may help impart curvature to the underlying beta sheet, contributing to the overall curvature of the traversing groove (Figure 13D).

Two additional conserved residues of the second patch, V13 and V16, do not participate in any hydrogen-bond interactions and have no obvious structural importance. Nevertheless, V16 likely has a critical, albeit unknown, role in protein function. V16 is analogous to the cathepsin C residue V249. Patients with V249F mutations develop Papillon-Lefèvre syndrome, again suggesting cathepsin C dysfunction.¹⁵⁴ Clearly, additional research is needed to further characterize these conserved patches and the traversing groove in which they are located (Figure 13D).

If we accept the hypothesis that *in vivo* the traversing groove constitutes a surface positioned at an important protein-protein interface, small molecules that target specific residues critical for protein binding may be able to disrupt the protein-protein interaction and potentially inhibit cruzain function. We note, however, that many of the residues that form the traversing groove are homologous with residues of human cathepsin C, and so relevant cruzain inhibitors are likely to inhibit cathepsin C as well. However, several residues, located between the two conserved patches (Figure

13B), are not themselves conserved. For example, the cathepsin C residues homologous to cruzain A15 and N47 are I258 and P224, respectively. It may therefore be possible to design cruzain-specific inhibitors that bind to non-conserved residues like A15 and N47.

Conclusion

Chagas disease, caused by the unicellular parasite *T. cruzi*, claims 50,000 lives annually¹¹ and is the leading cause of infectious myocarditis in the world.¹²⁹ As current antichagastic therapies like nifurtimox and benznidazole are highly toxic,^{16, 128, 130} ineffective at parasite eradication,¹³² and subject to increasing resistance,¹³¹ novel therapeutics are urgently needed.

Cruzain, the major cysteine protease of *T. cruzi*, is one attractive drug target.¹² In order to further the development of cruzain inhibitors, we here used MD simulations to identify a previously uncharacterized druggable pocket adjacent to the S2 subsite and a sequence alignment of a non-redundant, unbiased set of peptidase C1 family members to identify two conserved patches that may play roles in allosteric regulation, structural stability, or protein-protein interactions.

Future directions include using computer-aided drug design to identify and characterize cruzain inhibitors that exploit the previously uncharacterized pocket immediately beyond the S2 subsite. Considerably more effort is required to characterize and exploit the two conserved patches opposite the peptide-binding site. While several of the residues of these patches are known to be critical for the function

of cathepsin C, a cruzain homologue, mutagenesis studies are needed to directly confirm that they play an essential role in cruzain function as well. Once established as important, experiments are needed to further characterize the role these patches play. Two-hybrid screening or co-immunoprecipitation may identify other *T. cruzi* proteins that interact with cruzain. X-ray crystallography could then be used to determine whether or not these protein partners bind to the traversing groove formed by the two conserved patches identified in the current study. Additionally, small-molecule compounds that bind these patches may be useful tools for probing possible allosteric effects and/or disrupting critical protein-protein interactions.

5.4 Acknowledgements

Non-financial support from the National Center for Supercomputing Applications, the San Diego Supercomputer Center, the National Biomedical Computational Resource, and the Center for Theoretical Biological Physics is gratefully acknowledged. We would also like to thank Dr. Linda Brinen for helpful direction.

Chapter 5, in full, is a near reprint of the material as it appears in PLoS Neglected Tropical Diseases, volume 4, issue 5, e676. The dissertation author and Henrik Keränen were the primary co-contributors to this work. Benjamin A. Wilson also made significant contributions, and J. Andrew McCammon served as head adviser and principal investigator. PLoS journals have an open-access policy that permits republication of PLoS articles in academic theses.

CHAPTER 6: Toward Understanding the Conformational Dynamics of RNA Ligation

Robert V. Swift, Jacob Durrant, Rommie E. Amaro, and J. Andrew McCammon

6.1 Introduction

The intricate processing of genetic information is central to biological complexity and to the replication processes that sustain life. The discovery of post-transcriptional RNA editing was a rich addition to our appreciation of the central dogma of biology and is still a vibrant area of research with many unanswered questions. Within the mitochondria of kinetoplastids, a unique form of RNA editing exists whereby uridylates are inserted or deleted.¹⁵⁵ To carry out the editing process, kinetoplastids utilize a multi-protein complex known as the editosome. Though many details remain unresolved, the editosome is known to include a core 20S complex of 1.6 MDa containing 16-20 different proteins. Additionally, at least three distinct 20S complexes, distinguished by their constituent proteins and specificity for insertion or deletion editing, have been detected,¹⁵⁶ suggesting that an inherent dynamism in the 20S core may accommodate the varied specificity requirements of different gene products.

The editosome is localized in the mitochondria, where mitochondrial DNA has a topologically linked network of thousands of minicircles, which encode guide RNAs (gRNAs), and dozens of maxicircles, which encode components of respiratory complexes and energy-transduction systems. After transcription, gRNAs base pair with pre-mRNA from maxicircles *via* conserved “anchor sequences.”^{157, 158} Following gRNA base pairing, an endonucleolytic component of the editosome cleaves the pre-

mRNA at points of nucleotide mismatch between the pre-mRNA and the gRNA. Depending on the type of mismatch, uridylates are then inserted or removed from the nicked site of the pre-mRNA by a terminal uridylyl transferase or a U-specific 3' exonuclease, respectively.^{159, 160} One of two RNA-editing ligases, *TbREL1* or *TbREL2*, then religates the nicked RNA, generally depending on whether the preceding mRNA processing was insertion or deletion, respectively. The religated, “mature” mRNA is then ready for translation. *TbREL1* is relevant to structure-based rational drug design and associated structure-function studies both because its catalytically competent N-terminal adenylyltransferase domain has been crystallized in complex with ATP at 1.2 Å resolution⁸⁶ and because it is required for the survival of *T. brucei* insect and bloodstream forms.^{89, 161}

An ATP-dependent RNA ligase, *TbREL1* belongs to the nucleotidyltransferase superfamily, whose members include ATP and NAD⁺ dependent DNA ligases as well as eukaryotic mRNA capping enzymes.¹⁶² Each superfamily member shares five conserved active site structural motifs, along with a conserved catalytic mechanism in which a nucleotide is transferred to the 5' end of a polynucleotide *via* a covalent enzyme-NMP intermediate. Catalysis of dsRNA by *TbREL1* is initiated when a conserved catalytic lysine nucleophile attacks an ATP alpha phosphate, displacing pyrophosphate and forming an enzyme-AMP intermediate. In the second step, nicked dsRNA binds, and AMP is transferred from the enzyme to the 5'PO₄ end of the nick, forming a 5'-5' phosphoanhydride linkage. In the third step, the 3'OH end, which

apposes the 5'PO₄, attacks the phosphoanhydride linkage, displacing AMP and forming the religated dsRNA product.

Though the three religation steps of *TbREL1* are widely acknowledged, the exact dependence on divalent metal co-factors remains in doubt. For example, though the electron-density map of the ATP-bound *TbREL1* N-terminal domain clearly shows a Mg²⁺ coordinated between the non-bridging oxygen atoms of the ATP beta and gamma phosphates,⁸⁶ nucleotide transfer has not yet occurred, suggesting that a single Mg²⁺ may be insufficient to trigger the first chemical step, though the orthogonal position of the lysine nucleophile relative to the alpha-phosphate electrophile, which hinders nucleophilic attack, is an alternative explanation.¹⁶³ Additionally, kinetic evidence for related superfamily members suggests that the nucleotidyltransferase step proceeds via a two-Mg²⁺ mechanism,¹⁶⁴ while crystal-structure data from both the PBCV-1 capping enzyme and T4 RNA ligase 2 (Rnl2) show that a divalent metal co-factor in the vicinity of the alpha phosphate remains after step one of catalysis^{163, 165} and implies that a two metal mechanism may optimize the nucleotide transfer turn over rate. These contrasting results suggest that an allosteric signal, possibly from other protein members of the editosome, may lead to the uptake and coordination of a second Mg²⁺, thereby triggering nucleotide transfer.⁸⁶ The plausibility of ATP-dependent allosteric communication between *TbREL1* and other editosome members was demonstrated in molecular dynamics (MD) simulations of *apo* and ATP-bound *TbREL1*.⁸⁷ These simulations suggested that ATP binding induces increased mobility

of a putative allosteric loop; similar phenomena may provide the impetus for the uptake of a second Mg^{2+} .

Both sequence and structural phylogenetic analysis of superfamily members indicate that the closest known *TbREL1* homolog is the T4 phage Rnl2.^{86, 87} Recently, the structure of full-length Rnl2, in complex with its nucleic-acid substrate, was captured in two different conformations.¹⁶⁵ These two unique conformations illustrate the active-site conformational rearrangements that occur as the complex proceeds from the step-two product, immediately following step-two chemistry, to the step-three substrate, immediately preceding step-three chemistry. Additionally, comparing the crystal structure of human DNA ligase1 (LIG1), in complex with its DNA substrate, to the Rnl2 nucleic-acid complex reveals the salient features of substrate conformational requirements and substrate specificity. In both LIG1 and Rnl2, the nucleotides local to the nick are found in a RNA-like A-conformation. This conformation aligns the 3'OH and apposing 5'PO₄ at the nick site and is likely required for step-three chemistry. Normally in the canonical B-form, the DNA substrate of LIG1 is distorted into the required A-conformation *via* two phenylalanines that intercalate the minor groove. These phenylalanines, conserved in DNA ligases, are absent in RNA ligases like Rnl2 and *TbREL1*. Further details are presented in the original crystal structure paper¹⁶⁵ and are reviewed elsewhere in the context of other DNA and RNA ligases.¹⁶⁶

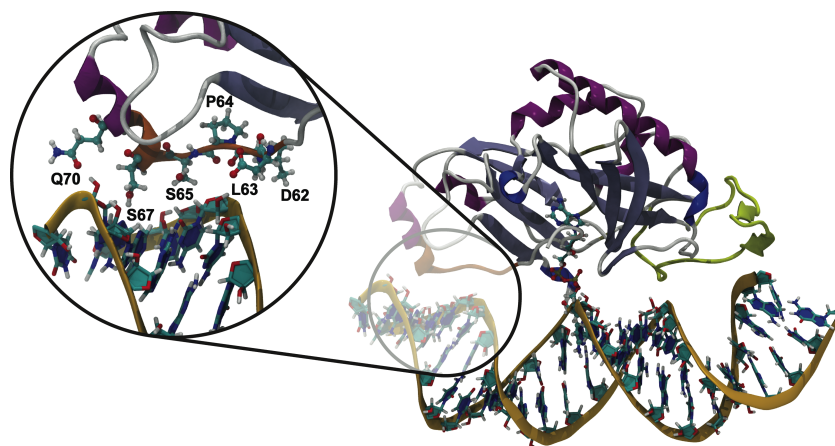


Figure 14: Nicked dsRNA bound to *TbREL1*, poised to catalyze religation. A unique kinetoplastid insert is colored in yellow, and a cluster of residues thought to be important in RNA recognition on the 5'PO₄ side of the nick is shown in orange. The magnified insert shows protein residues that make stabilizing contacts with the RNA.

Motivated by the close homology between Rnl2 and *TbREL1* and the desire to understand the functional dynamics of the kinetoplastid ligation reaction, we investigate the conformational dynamics of *TbREL1* in complex with a nicked dsRNA substrate in its canonical A-form. Based on the homologous structural and functional relationships between *TbREL1* and T4 Rnl2, we develop a model of a nicked dsRNA-*TbREL1* complex (Figure 14) and perform an all-atom explicitly solvated molecular dynamics simulation of the system. The integration of RNA into MD simulations has benefited from recent improvements in both force-field development and algorithms,^{167, 168} allowing us to make insights into the dynamics and catalytic mechanism. The simulated system contains AMP attached to the nick 5'PO₄ moiety (AppN), characteristic of the step-two intermediate.¹⁶⁵ The non-nicked strand represents the template, or gRNA, and the nicked strand represents the post-

transcriptionally modified pre-mRNA. Fluctuations of conserved interactions between the enzyme and RNA substrate, as well as the dynamics of conserved active-site residues, are examined and discussed in the context of step three of the religation mechanism. Furthermore, we explore the effect of a unique kinetoplastid insert region on dsRNA binding.

6.2 Materials and Methods

System preparation. A model of the step-two intermediate of *TbREL1* in complex with A-form dsRNA was built in homology to chain B of the T4 Rnl2 crystal structure (PDBID 2HVR), which captures T4 Rnl2 immediately preceding step-three chemistry.¹⁶⁵ A 12-mer of A-form dsRNA, with a sequence identical to the nicked DNA/RNA hybrid co-crystallized with T4 Rnl2, was generated using the NUCGEN module of Amber 9,¹⁶⁹ which creates accurate nucleotide models based on fiber-diffraction studies. The dsRNA was then aligned to the nicked DNA/RNA hybrid by minimizing the root-mean-square-deviation (RMSD) between the two nitrogenous bases of the DNA/RNA hybrid that flanked the nick and the commensurate nitrogenous bases of the model dsRNA. Next, the dsRNA was translated onto the crystal structure of the *TbREL1* N-terminus (PDBID 1XDN) by minimizing the RMSD between the AppN adenine moiety of the DNA/RNA hybrid and the adenine moiety of the ATP bound in the *TbREL1* active site. A nick was then created in the model dsRNA identical to the nick of the DNA/RNA hybrid bound to T4 Rnl2. The beta and gamma phosphates of the *TbREL1*-bound ATP were subsequently deleted,

and the O-P_{AMP}-O-P_{5'PO4} dihedral angle was brought into agreement with the corresponding dihedral angle of the T4 Rnl2 crystal structure. The resulting AMP is linked through its phosphate to the 5'PO₄ nicked terminal nucleotide; we designate this AMP-phosphate-phosphate-nucleotide residue AppN according to the nomenclature of Nandakumar and co-workers.

To ensure that the active site of our model reflected the conformation of Rnl2 immediately preceding step-three chemistry, the *TbREL1* active-site residue R111 was given a conformation commensurate with its Rnl2 homologue, R55. Specifically, R111 was rotated away from the non-bridging oxygen of the AppN alpha-phosphate and placed within hydrogen-bonding distance of the Y58 backbone carbonyl. The conformations of the other *TbREL1* active-site residues were in close agreement with the conformations of the homologous active-site residues of T4 Rnl2 immediately preceding step-three chemistry, and so required no further manipulation. A Mg²⁺ was manually positioned approximately 2 Å from a non-bridging AMP alpha phosphate based on the assumptions that ion dissociation occurs neither during RNA binding nor during subsequent active-site rearrangements and that the *TbREL1* catalytic mechanism proceeds via a two-Mg²⁺ mechanism. This Mg²⁺ position was chosen based on the crystal-structure data of both the PBCV-1 capping enzyme¹⁶³ and the T4 Rnl2 step-one product,¹⁶⁵ as well as on a similarly positioned Ca²⁺ in the crystal structure of T4 Rnl1 bound to the unreactive analog AMPCPP.¹⁷⁰ Five additional water molecules, each approximately 2 Å from the Mg²⁺ ion, were manually positioned to duplicate the expected octahedral coordination complex.

All co-crystallized water molecules in the *Tb*REL1 structure were retained in our model. Selenomethionines used for crystal structure refinement were replaced with methionines, and WHATIF¹⁷¹ was used to assign histidine protonation states. Protonation-state assignments were manually verified. With the exception of the non-standard AppN residue, all RNA and protein hydrogens and force-field parameters were assigned according to the Amber ff99SB force field¹³⁷ using the xleap module in Amber 9.¹⁶⁹ The non-standard residue AppN is composed of an adenosine linked to a cytidine *via* a phosphoanhydride linkage; consequently, we assigned AppN charges using the RA5 and RC5 residues of the Amber ff99SB force field as a template. The charge of the oxygen bridging the phosphoanhydride linkage was assumed to be identical to that of the oxygen bridging the alpha and beta phosphates in the Carlson ADP parameter set.¹⁷² All missing AppN force-field parameters came from the GAFF force field.¹³⁶ Following parameterization, the complex was immersed in a rectangular box of TIP3P waters⁵⁵ with a 10 Å buffer from the protein to the edge of the water box in each direction. Sodium and chloride ions were added to the system, bringing it to electric neutrality and yielding a 100 mM salt solution.

MD simulations. Spurious contacts were removed through a four-tier set of 45,000 energy minimization steps. The first 10,000 steps were performed in two 5,000-step cycles. Hydrogens were relaxed during the first 5,000 steps, holding all other atoms fixed. Hydrogens, waters, and ions were relaxed during the next 5,000 steps, holding the protein and nucleic acid fixed. Ten thousand steps of minimization

followed in which only the protein backbone was fixed. No constraints were applied during the last 25,000 steps of minimization.

To equilibrate our model, we performed a one-nanosecond restrained simulation using the minimized system as a starting point. Temperature was introduced at 310 K with harmonic restraints on the protein backbone. Through the gradual relaxation of the backbone restraints, the protein structure was preserved as kinetic energy was introduced. For the first 250 ps, a harmonic restraining force of 4.0 kcal/mol/Å² was applied to the backbone atoms. Similarly, for the second, third, and fourth 250 ps segments, the harmonic restraints were consistently weakened to 3.0, 2.0, and 1.0 kcal/mol/Å², respectively.

A 70 ns MD simulation followed in which unrestrained dynamics were performed with a one-fs timestep. The duration of the simulation was sufficient to equilibrate the system based on RMSD values (overall alpha carbon [Fig SI1] and the 3'-terminus sugar pucker phase angle [Figure 16]), while allowing us to sample functionally relevant conformations local to the step-2 intermediate. The temperature bath was maintained by Langevin dynamics at 310 K, and pressure was maintained with the hybrid Nose Hoover-Langevin piston method at 1 atm⁵⁶ using period and decay times of 100 and 50 fs, respectively. The Particle Mesh Ewald algorithm was used to treat long-range electrostatics.⁵⁷ A multiple time-stepping algorithm was employed where bonded interactions were computed at every time step, short-range non-bonded interactions were computed every two time steps, and full electrostatics were computed every two time steps. The water hydrogen-oxygen and hydrogen-

hydrogen distances were constrained using the SHAKEH algorithm⁵⁸ to be within 0.0005 of the nominal lengths. All minimization and MD simulations were carried out with NAMD 2.6⁵⁹ using the Amber ff99SB force field.¹³⁷ Simulations were performed on the TeraGrid ABE cluster. A typical benchmark for the 57,575-atom system on 128 processors was 0.21 days per nanosecond of simulation. System configurations were sampled every 50 ps, generating 1,400 coordinate snapshots for subsequent analysis.

Pucker analysis The pseudorotational formalism proposed by Rao, Westhof, and Sundaralingam¹⁷³ was used to characterize the puckering of the nick 3' ribose ring. We first defined the following Fourier sums:

$$A = \frac{2}{5} \sum \theta_j \cos\left(\frac{4\pi j}{5}\right)$$

$$B = -\frac{2}{5} \sum \theta_j \sin\left(\frac{4\pi j}{5}\right)$$

where j ranges from zero to four, and $\theta_0, \theta_1, \theta_2, \theta_3, \theta_4$ are the torsion angles about C2'—C3', C3'—C4', C4'—O4', O4'—C1', and C1'—C2', respectively. The pucker phase angle, P , identifies which part of the ring is furthest from planarity. Ribose configurations generally take on the 2' endo pucker typical of DNA or the 3' endo pucker typical of RNA. The possible values of P , which range from zero to 360, correspond to the stages along the so-called pseudorotation pathway. The ribose ring has five atoms, each of which can be above or below the plane, yielding 10 distinct stages. The envelope conformations correspond to $P = 0^\circ, 36^\circ, 72^\circ$, etc. and alternate

with the twist conformations at $P = 18^\circ, 54^\circ, 90^\circ$, etc. In the Rao-Westhof-Sundaralingam pseudorotational formalism,

$$P = \tan^{-1}\left(\frac{B}{A}\right)$$

The maximal torsion angle or puckering amplitude, θ_m , measures the magnitude of the deviation from planarity.¹⁷³

$$\theta_m = \sqrt{A^2 + B^2}$$

Polar plots are generally used to represent sugar pucker, with the puckering amplitude and phase angle assigned as the radial and angular coordinate, respectively. Analysis of experimental structures shows that the 2' endo ribose configuration typical of DNA has $137^\circ < P < 194^\circ$, and the 3' endo ribose configuration typical of RNA has $-1^\circ < P < 34^\circ$.¹⁷⁴

Principal Component Analysis. Principal component analysis, or PCA, is a common statistical data-analysis technique whose application to the field of biomolecular simulation was pioneered by Berendsen¹⁷⁵ and Garcia.¹⁷⁶ By diagonalizing a covariance matrix constructed of molecular coordinates sampled over the duration of a MD trajectory and listing the resulting eigenvectors in order of descending eigenvalue magnitude, PCA separates large, low frequency conformational changes from small, high frequency motions. By considering only the conformational fluctuations of the first few modes, researchers frequently use PCA to filter out large isomerization events and to reduce the complexity of trajectory analysis. In this work,

PCA was used to explore the large, low frequency motions of the dsRNA substrate. Following trajectory equilibration, principal component analysis was carried out on the phosphate atoms of the dsRNA substrate, using the ptraj module in Amber 9.¹⁶⁹ Prior to performing PCA, translational and rotational motions were removed by fitting the sampled coordinates to the first equilibrated frame of the trajectory.

6.3 Results and Discussion

Equilibration. A 70 ns constraint-free simulation was performed on the dsRNA-bound *TbREL1* system to examine *TbREL1* dynamics. A plot of the time evolution of the RMSD of the protein alpha carbons, in which each trajectory frame was aligned to the initial frame to remove rotational and translational motion, suggested that our preliminary efforts to equilibrate the system prior to our 70 ns simulation were successful. However, subsequent examination of the distance between the nick 3'OH and the phosphorus atom of the 5'-phosphoanhydride linkage suggested that equilibration was not complete until after 38.75 ns. During the initial phase (0-38.75 ns), these two atoms, between which a bond forms following step three, were distant ($7.16 \text{ \AA} \pm 1.20$). During the second phase (38.75 ns to 70 ns), however, the nick 3'OH and the nick 5' phosphorus moved closer together ($3.95 \text{ \AA} \pm 0.29$), as required for catalysis. Correspondingly, the trajectory time from 0 to 38.75 ns is referred to as the “equilibration phase,” and from 38.75 to 70 ns as the “dynamics phase.” Unless otherwise specified, all analysis was performed on the dynamics phase.

Active-Site Interactions. The atomic interactions within the active site can be divided into three categories: interactions between *Tb*REL1 and the adenine moiety of AppN, interactions between *Tb*REL1 and the ribofuranose of AppN, and interactions between *Tb*REL1 and the pyrophosphate of AppN. Throughout our discussion, we adopt the naming scheme of Nandakumar and coworkers, naming the nicked strands based on their active-site ends (i.e. the 5'PO₄ strand and the 3'OH strand).¹⁶⁵

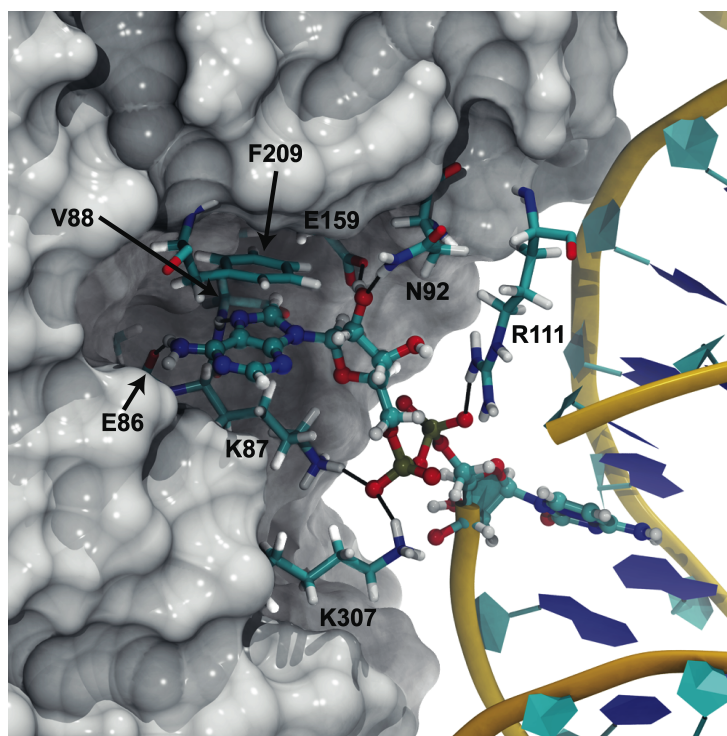


Figure 15: The equilibrated structure illustrates the close contacts between key active-site residues (shown in licorice), the AppRNA (shown in ball and stick), and the 3'OH RNA strand.

Table 6: Active-site interactions, including protein-AppN interactions, protein-nick interactions, and interactions with the Mg^{2+} -coordinated water molecules.

	<i>Length [$\text{\AA}$] (Std) [$\text{\AA}$]</i>
AppN adenine – F209 phenyl group	4.14 (0.22)
E86:O – AppN adenine N6	2.97 (0.20)
E159:OE1 – AppN ribose O2'	2.55 (0.08)
N92:ND2 – AppN ribose O2'	3.30 (0.41)
K87:NZ – AppN:O1P	3.39 (0.42)
K307:NZ – AppN:O2P	3.87 (0.56)
R111:NH1 – AppN:O1P	2.80 (0.11)
Nick 3' HO'3 – nearest non-bridging nick 5'PO ₄ oxygen	1.90 (0.64)
R111:CZ - nick O3'	3.95 (0.34)
H89:ND1 – WAT:O	2.85 (0.21)
E159:CD – WAT:O	3.49 (0.27)
E283:CD – WAT:O	3.45 (0.24)
E283:CD – WAT:O	3.44 (0.21)
nick O2' – WAT:O	3.30 (0.53)

Deep within the active site, F209, E86, and V88 interact with the adenine moiety of AppN and stabilize its motion (Figure 15). F209 acts as a conserved hydrophobic platform that forms pi-pi interactions with the aromatic adenine moiety of AppN, preventing large lateral fluctuations of this base within the binding pocket. Hydrophobic stabilization in this position is essential for step-three catalysis in other type II RNA ligases, as demonstrated by the partial recovery of religation activity when a leucine residue was substituted for the F209 homolog in an alanine scan of T4 Rnl2.¹⁷⁷ Our simulation is consistent with this mutagenesis study; F209 exhibits minor fluctuations, with an average distance of 4.14 Å between the centers of mass of its phenyl group and the AppN adenine moiety (Table 6). The movement of AppN is also restrained by E86, which accepts a hydrogen bond from the adenine N6-amine that holds the E86 O1 (or O2) oxygen within 2.97 Å, on average (Table 6). Conserved among type II RNA ligases, E86 is essential in mediating step-three catalysis in T4

Rnl2.¹⁷⁷ The observed distances and fluctuations of both F209 and E86 are similar to those observed in a previous simulation of the ATP-*TbREL1* complex.⁸⁷ These results, along with the T4 Rnl2 mutational analysis,¹⁷⁷ reflect the importance of F209 and E86 in mediating the re-ligation of the adenylated dsRNA substrate and the formation of the adenylated enzyme intermediate.

Stabilizing interactions between the protein and the AppN ribofuranose complement the contacts formed between the adenine moiety and residues deep within the binding pocket. In particular, E159, conserved across all members of the superfamily, and N92, conserved across all type II RNA editing ligases,¹⁷⁷ make critical hydrogen bonds with the ribofuranose O2' hydroxyl group (Figure 15, Table 6). The close, stable contact made by E159 was observed in a previous simulation of the ATP-bound complex⁸⁷ and is consistent with the conservative mutational analysis, which identified E159 as an essential mediator of steps one and three of catalysis.¹⁷⁷ Similarly, an alanine scan of T4 Rnl2 revealed that the N92 homolog plays a critical role in the religation of the pre-adenylated dsRNA substrate.¹⁷⁷ No isomerization of the AppN ribofuranose moiety was observed during the course of catalysis in the T4 Rnl2 homolog¹⁶⁵; consequently, the mutational analysis results, coupled with the short, stable contacts observed during our simulation, suggest that the N92-ribofuranose interaction helps maintain the ribofuranose conformation throughout the catalytic cycle.

Charged residues at the periphery of the binding site, K87, K307, and R111, each interact with the AppN pyrophosphate and stabilize the step-three substrate

conformation (Figure 15). The orthogonal orientation of the K87 nucleophile to the alpha phosphate is typical of the step-three substrate conformation and persists throughout the 31.25 ns dynamics phase. The system does not undergo isomerization away from the step-three substrate conformation, indicating that the network of interactions captured in the homologous T4 Rnl2 crystal structure is a reasonable model of the *TbREL1* active-site architecture prior to step three of catalysis.

Interestingly, K87 remains within hydrogen-bonding distance of a non-bridging pyrophosphate oxygen during the dynamics phase of the simulation (Table 6). This result suggests that the K87 nucleophile helps maintain the protein in the step-three conformation once isomerization into that conformation has occurred. Consequently, K87 seems to perform a dual role during catalysis: during step one it plays the central chemical role as the nucleophile, and during step three it plays a peripheral stabilizing role as a hydrogen-bond donor. To the best of our knowledge, no experimental evidence has been published supporting this hypothesis, but the putative dual role of K87 could be probed using the methods Yin and co-workers employed during their structure-function analysis of T4 Rnl2.¹⁷⁷ Nearby K307 also donates a hydrogen bond to a non-bridging pyrophosphate oxygen, remaining close throughout the dynamics phase of the simulation (Table 6). The AppN stabilization that this interaction provides is in agreement with mutational analysis, which has shown that a positively charged hydrogen-bond donor at this position is most efficient for religating preadenylated nicked-dsRNA. Along with K87 and K307, R111 also donates a

hydrogen bond to one of the non-bridging pyrophosphate oxygens and remains in close contact throughout the dynamics phase of the simulation (Table 6).

Insights into catalysis. A requirement for step-three catalysis is that the O3' nucleophile and the departing, bridging oxygen of the AppN pyrophosphate take on an apical, or “in-line” arrangement. This apical requirement is achieved only when the terminal ribofuranose of the nucleotide on the 3'OH side of the nick takes on a 3' endo conformation.¹⁷⁸ The 3' endo conformation is the native conformation of dsRNA, explaining why T4 Rnl2 requires that the strand on the 3'OH side of the nick be RNA while showing no preference for nucleic-acid type on the 5'PO₄ side.¹⁷⁸ *In vitro* religation experiments with *Tb*REL1 alone have likewise demonstrated that the strand on the 3'OH side of the nick must be RNA,¹⁷⁹ though inclusion of other components of the editosome complex alters substrate specificity. Here, we assume based on the close homology between *Tb*REL1 and T4 Rnl2 that the RNA requirement on the 3'OH side of the nick arises from the same 3' endo requirement observed in T4 Rnl2. The equilibrium distribution of sugar conformations observed during our simulation should be a reasonable approximation of those available for catalysis during step-three religation.

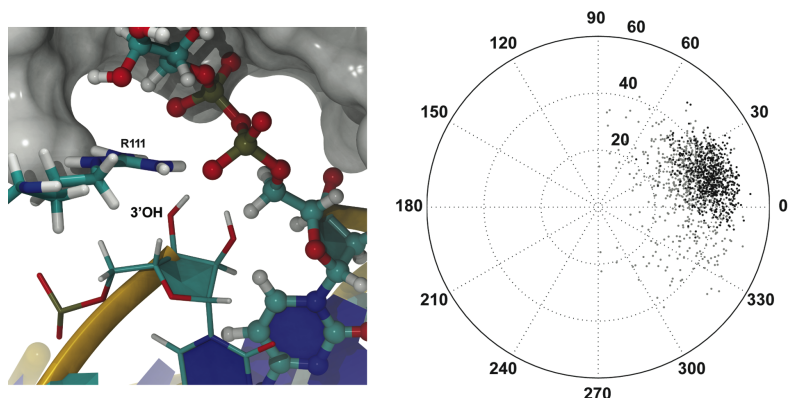


Figure 16: The equilibrated ribofuranose pseudorotation angle for the 3'OH nucleotide as depicted within the active site (left panel) and in polar coordinates (right panel, equilibration data in gray, dynamics data in black).

In order to gain insight into the temporal fluctuations of the required 3' endo conformation, we monitored the ribofuranose pseudorotation angle of the nucleotide at the 3'OH terminal of the nick over the dynamics phase of the simulation (Figure 16). The pseudorotation angle is a measure of the sugar conformational space and includes two stable conformations corresponding to the 2' endo and 3' endo conformations of DNA and RNA, respectively. As noted before, the 2' endo conformation is characterized by a pseudorotation angle between 137° and 194° , and the 3' endo conformation is characterized by a pseudorotation angle between -1° and 34° .¹⁷⁴ Consistent with the A-form dsRNA structure, the ribofuranose of the nick terminal 3' hydroxyl nucleotide remains almost exclusively in the 3' endo conformation ($15.2^\circ \pm 10.9^\circ$).

Visual inspection of the O3' nucleophile and the bridging oxygen leaving group of the AppN pyrophosphate demonstrated that while the two atoms remain parallel to one another, they are not strictly in-line; the attacking O3' nucleophile is vertically displaced from the plane of the leaving group (Figure 16). The vertical

displacement is a consequence of relaxing steric conflicts in the initial model during minimization and equilibration and implies that the active-site conformation is sub-optimal for religation. Despite this, the time evolution of active-site topography still provides a useful model from which mechanistic insight may be inferred.

The nicked 3' hydroxyl nucleophile must be deprotonated either prior to, or following, reaction with the beta phosphate of the AppN intermediate. Determining the lowest-energy pathway of proton transfer is complicated and highly dependent on the electrostatic environment surrounding the acid. However, proton transfer is restrained by the requirement that a new covalent bond forms between the acidic hydrogen and the base. This requirement allows simple hypotheses of proton transfer pathways to be made based on the proximity of the acidic hydrogen to potential general bases. Our simulation demonstrates that the acidic hydrogen of the nicked 3' hydroxyl nucleophile resides at an average distance of 1.90 Å from the nearest negatively charged non-bridging pyrophosphate oxygen (Table 6). Furthermore, protein fluctuations bring these two atoms within 1.40 Å of each other, causing significant overlap of their van der Waals radii and suggesting possible proton transfer between these two atoms. As the acidic hydrogen lies between the putative basic oxygen and the nucleophilic O3', it is likely that proton transfer either precedes nucleophilic attack or occurs concertedly with phosphodiester formation. Additionally, the positively charged guanidinium moiety of R111 is proximal to the nick O3' hydroxyl (Table 6) and may play a role in stabilizing the O3' as it acquires negative charge during the proton-transfer process. This is consistent with mutational

analysis carried out on the T4 Rnl2, which demonstrated that a lysine mutant in the position of the R111 homolog was nearly as efficient in religating preadenylated nicked dsRNA as the wild type enzyme, but that the alanine mutant lacked religation ability.¹⁷⁷ The idea that a non-bridging oxygen of a nucleotide phosphate may serve as a general base has also been proposed as a step in the catalytic mechanism of the PBCV-1 mRNA capping enzyme.¹⁶³

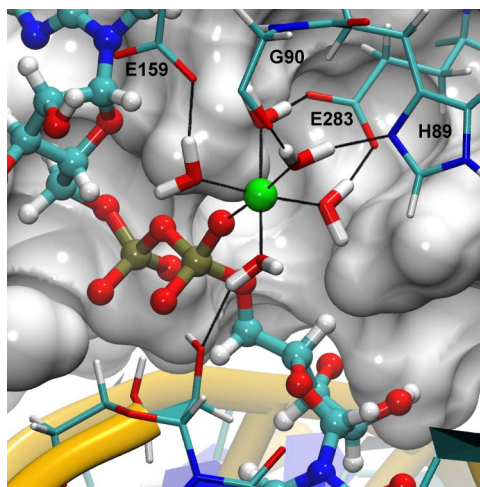


Figure 17: Equilibrated coordination of the active-site magnesium ion.

Mg²⁺ coordination. The exact dependence of nicked dsRNA religation on divalent metal cofactors remains uncertain. In our model of the step-three substrate, we assume that *TbREL1* step-one chemistry proceeds via a two Mg^{2+} cofactor mechanism. This assumption, along with homology to the crystal structures of other nucleotidyltransferase superfamily members, suggests the Mg^{2+} placement in our model (Figure 17). The first Mg^{2+} cation, which is initially coordinated by the non-bridging oxygens of the AppN beta and gamma phosphates, dissociates with pyrophosphate following the formation of the enzyme nucleotide intermediate in step

one. The second Mg^{2+} cation, which is coordinated in the vicinity of the alpha phosphate, remains bound within the active site to mediate later steps of catalysis. With the model so developed, analysis of the network of interactions mediated by the presence of the divalent magnesium, as well as the proximity of the cation to the AppN substrate, suggest possible functional roles for the bound metal in the context of step-three chemistry.

In our model, the magnesium cation is octahedrally coordinated by five water molecules and a non-bridging alpha-phosphate oxygen of the AppN pyrophosphate (Figure 17). None of these water molecules exchanged with bulk solvent during the simulation. Four of the five water molecules participate in a hydrogen bond network with conserved residues in motifs I, III and IV, while the fifth water donates a hydrogen bond to the 2' oxygen of the terminal nucleotide on the nicked 3' hydroxyl strand. This interaction network provides bridging hydrogen bond interactions to conserved active-site residues and serves to maintain the relative position of the magnesium within the active site, which may be important in transition-state stabilization as the reaction proceeds.

In phosphoryl transfer reactions, there are three mechanistically distinct nucleophilic substitution pathways: dissociative, concerted, and addition-elimination.¹⁸⁰ The addition-elimination and concerted pathways are $\text{S}_{\text{N}}2$ -like; as the nucleophile approaches the phosphate electrophile, the reacting phosphate passes through a planar, pentacoordinate transition state in which one of the non-bridging oxygens acquires an additional negative charge.^{180, 181} In the dissociative, E_{I} -like

pathway, the leaving group first departs, forming a planar metaphosphate, phosphocation intermediate before the nucleophile reacts. In our model of the step-three substrate, the Mg^{2+} cation coordinates one of the non-bridging oxygens of the phosphate electrophile. The magnesium resides at an average distance of 1.83 Å from the non-bridging oxygen, opposite the O3' nucleophile across the pyrophosphate backbone (Figure 17). The proximity of the divalent magnesium and the non-bridging oxygen of the phosphate electrophile may provide electrostatic stabilization of the excess transition-state negative charge typical of the concerted and addition-elimination pathways, enhancing the turnover rate. Conversely, if step-three chemistry proceeds via a dissociative pathway, the proximity of the positively charged divalent magnesium would destabilize the phosphocation of the metaphosphate intermediate, thereby attenuating the turnover rate. While these considerations are qualitative and non-rigorous, they suggest that the presence of a divalent metal during step-three chemistry favors either a concerted or an addition-elimination pathway.

Having considered the effects of a divalent metal on the turnover rate of the third chemical step, we now consider the network of interactions provided by conserved residues, which stabilize the magnesium within the active site. Within motif I, both the unprotonated delta nitrogen of H89 and the carbonyl oxygen of G90 accept hydrogen bonds from two unique water molecules, each coordinating Mg^{2+} (Figure 17, Table 6). Within motif III, one of the negatively charged carboxyl oxygens of E159 accepts a hydrogen bond from a coordinating water molecule. The carboxyl oxygens of E283 in motif IV likewise accept hydrogen bonds from two

coordinating water molecules in a similar manner (Table 6). When the E159 or E283 homologs are substituted with either aspartate or glutamine in T4 Rnl2, religation activity is completely abolished, indicating a strict glutamate requirement in these positions.¹⁷⁷ Although the cause of this dependence is not immediately apparent, we speculate that neither glutamine, which is neutrally charged, nor aspartate, whose side chain is one methylene group shorter, provide stable hydrogen-bonding partners for the magnesium-coordinating waters.

In addition to the E159 and E283 interactions, one of the Mg^{2+} -coordinating waters also donates a hydrogen bond to the 2' oxygen of the terminal nucleotide on the nicked 3' hydroxyl strand (Figure 17, Table 6). This interaction indirectly links the terminal nucleotide of the nicked 3' hydroxyl strand to the AppN moiety and may play a role in properly orienting the terminal nucleotide for step-three chemistry.

RNA-Binding Footprint. While the nucleotide specificity determinants for the T4 Rnl2 and human DNA ligase I superfamily members have been clearly defined,¹⁶⁶ the specificity determinants of *TbREL1* have not. Available experimental data demonstrates that, *in vitro*, *TbREL1* cannot religate nicked nucleic acid in complex with gRNA unless both of the constituent fragments are RNA.¹⁷⁹ Based on studies of the homologous T4 Rnl2, we hypothesize that the RNA requirement on the 3'OH side of the nick may arise at least in part from the need to confer a 3' endo conformation at the 3'OH nicked terminal nucleotide.¹⁷⁸ The absence of religation when DNA is present on the 5' side of the nick suggests that specific *TbREL1*-RNA interactions play

important roles in mediating substrate stability during steps two and three of the religation mechanism.

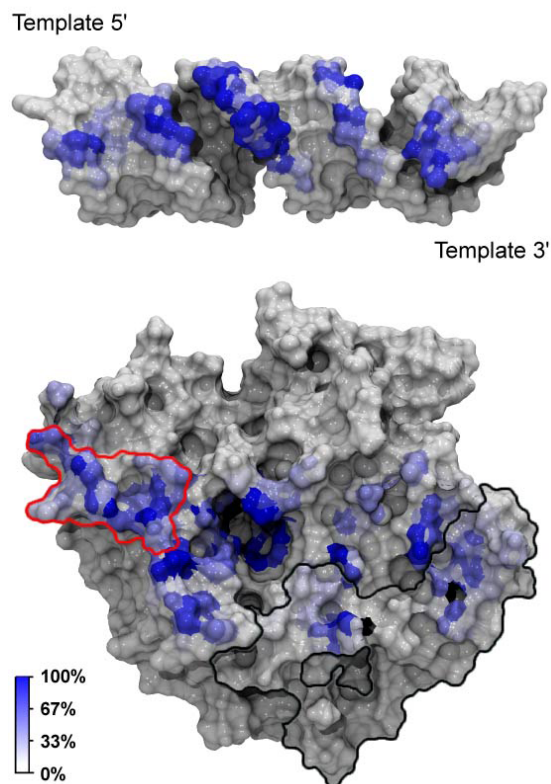


Figure 18: RNA binding footprint. Top panel: the protein dynamic contact footprint projected onto the molecular surface of the RNA. Bottom panel: the RNA dynamic contact footprint projected onto the molecular surface of *TbREL1*. Both the protein and RNA are shown in gray surface representation with the contacts colored by percent occupancy over the dynamics portion of the simulation. A unique kinetoplastid insert is outlined in black, and a cluster of residues thought to be important in RNA recognition on the 5'PO₄ side of the nick is outlined in red.

In order to identify molecular specificity determinants that mediate dsRNA binding, we generated a RNA binding footprint based on a set of protein-RNA interaction persistence times (Figure 18). This analysis allows us to identify primary contact sites between the RNA and the protein and to discriminate longer-lived interactions from more transient ones. The footprint-like mapping of the protein-RNA

interaction persistence times onto the protein and RNA surfaces shows that more extensive interactions occur on the 5'PO₄ side of the nick. Although this contact pattern contrasts with the results of Nandakumar and co-workers, who report that the closest homolog of *TbREL1*, T4 Rnl2, makes more extensive contacts with the strands on the 3'OH side of the nick, it supports the experimental observations that *TbREL1* requires RNA on the 5'PO₄ strand of the nick for *in vitro* polynucleotide religation. A large, contiguous set of interactions, as well as various individual interactions, makes up the more extensive contacts on the 5'PO₄ side of the nick. Specifically, D62, L63, P64, S65, S67, and Q70 collectively form a contiguous surface, herein called the “recognition sequence,” which intercalates the minor groove of the dsRNA on the 5'PO₄ side of the nick (Figures 14 and 18). Although the recognition sequence is not conserved at the superfamily level, it is conserved among the Trypanosomatidae type I mitochondrial RNA-editing ligase sequences available in the SWISS-PROT database.¹⁸² As the depth and width of the minor grooves differ between the canonical RNA and DNA forms, this recognition sequence may play a key role in polynucleotide discrimination.

Table 7: RNA-protein interactions. Distances and occupancy percentages are calculated over the dynamics phase only.

<i>Protein</i>	<i>RNA</i>	<i>Length [$\text{\AA}$] (Std) [$\text{\AA}$]</i>	<i>% Time Bond < 3.5 $\text{\AA}$</i>
LEU63:N	C _g -7:O1P	3.84 (0.99)	55.50%
GLN70:NE2	U _n -11:O1P	5.34 (1.78)	19.62%
SER67:OG	A _n -10:N3	4.67 (1.29)	24.4%
SER65:OG	C _g -8:O2'	4.86 (1.43)	18.34%
PRO64:O	U _g -9:O2'	4.14 (0.68)	19.62%
ASP62:OD1	C _g -8:O2'	5.49 (1.64)	20.73%
MET314:O	A _n -1:O2'	3.57 (1.22)	71.29%
GLU119:OE1	C _n +12:O5'	6.37 (4.17)	25.84%
ASN120:ND2	A _n +11:O1P	3.36 (1.74)	88.20%
PHE121*	U _g +4:O2'	7.85 (0.99)	0.00%
PHE122*	G _g -3:O2'	5.19 (0.48)	0.00%
ARG194:NH2	G _g +3:O1P	3.09 (0.45)	82.46%
ARG309:NH1	A _n -1:O1P	2.82 (0.12)	99.84%
ARG309:NH1	C _n -2:O2P	2.94 (0.33)	94.90%

*Side-chain ring center of mass.

In addition to the geometric complementarity conferred by these residues, several specific atomic interactions within this sequence may contribute to polynucleotide discrimination (Table 7). The carboxyl group of D62, the carbonyl backbone of P64, the side-chain hydroxyl of S65, and the side-chain amino group of Q70 each make contact with a ribofuransose 2'OH. Two of these four RNA-specific contacts are shared between D62 and S65; roughly nineteen nanoseconds into the dynamics portion of the simulation, the 2'OH interaction with S65 exchanges with D62 and persists throughout the remainder of the simulation. In DNA, the lack of the 2'OH precludes hydrogen-bond interactions with D62, P64, S65, and Q70. These interactions may partly explain why the nucleic acid on the 5'PO₄ side of the nick must be RNA for religation to occur *in vitro*.

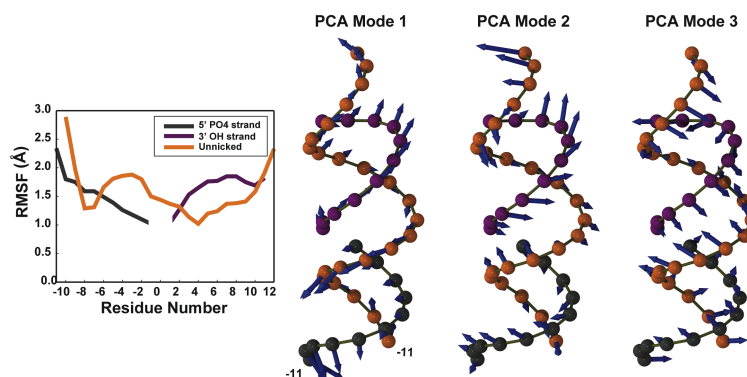


Figure 19: RNA flexibility as shown by both the RMSF per residue of the RNA strands and PCA analysis. In both panels, the template strand is shown in orange, the 3'-OH strand in gray, and the 5'-PO₄ strand in purple. PCA vectors were multiplied by a factor of 25 for easier visualization.

The stability conferred by the recognition sequence is immediately apparent upon examining the root-mean-square-fluctuation (RMSF) values and the porcupine plots describing the motion along the first three principal modes (Figure 19). The recognition sequence makes contact with the unnicked gRNA strand at positions -8 and -7, causing a two-fold decrease in mean-square fluctuation relative to that observed at the 5'PO₄ terminal of the gRNA (position -10). At positions -7 to 2, where the gRNA nucleotides are not in contact with the protein, there is a 1.6 fold decrease in mean-square fluctuation. Additionally, the porcupine plots illustrate that the magnitudes of atomic displacements along the first three modes are attenuated in positions -8 and -7 (Figure 19). The relatively low interaction persistence times observed for the recognition sequence (Table 7) are readily explained by the relatively large positional fluctuations of the dsRNA at the terminal ends on the 5'PO₄ side of the nick (Figure 19).

Two additional independent *Tb*REL1-RNA interactions play a role in mediating substrate stability on the 5'PO₄ side of the nick. The carbonyl of M314

forms a hydrogen bond with the 2'OH of the penultimate nucleotide of the nicked strand, contributing to RNA specificity (Table 7). Additionally, the guanidinium moiety of R309 bridges the penultimate and antepenultimate RNA residues on the 5'PO₄ side of the nick by forming two stable hydrogen bonds with two non-bridging phosphate oxygens. The lack of RNA-specific interactions with R309 suggests that this residue may confer stability on the bound state without affecting substrate specificity.

Aside from the hypothesis that the 3'OH nucleic acid adjacent to the nick must take on the RNA-like 3' endo conformation for catalysis to proceed, the obligate presence of RNA on the 3'OH side of the nick may also stem from specific RNA-protein interactions that possibly impart stability to the bound complex. The 2'OH of the terminal nucleotide on the 3'OH side of the nick interacts with the carboxyl group of E119 and attenuates fluctuations of the terminal dsRNA, thereby conferring stability to the bound complex (Table 7). N120 also provides stabilizing interactions with the RNA phosphate backbone contributing to RNA stability, though probably not specificity (Table 7). Two phenylalanine residues, F121, F122 contribute non-specific van der Waals interactions with adjacent ribonucleotides in the gRNA strand on the 3'OH side of the nick.

Mutational religation assays carried out on T4 Rnl2 suggest important roles for both F121 and F122; the F121 homolog is an important contributor to non-specific polynucleotide stabilizing interactions, and the F122 homolog is important in conferring RNA specificity. Additionally, the DNA/RNA-T4 Rnl2 crystal structure

shows both the F121 and F122 homologs intercalating the minor groove.¹⁶⁵ In light of the experimentally determined importance of the F121 and F122 homologs and their crystallographically determined positions, which our initial homology model mimicked, it is appropriate to explore why F121 and F122 were not found intercalating the minor groove during the equilibrated, dynamics phase of the trajectory. Roughly 21 ns into the equilibration phase of the simulation, the dsRNA on the 3'OH side of the nick moved away from the protein surface. When it returned, approximately 2 ns later, F121 and F122 no longer penetrated the minor groove. Instead, they made non-specific van der Waals contacts with the two adjacent ribofuranose residues that border the 3'OH side of the minor groove in which F121 and F122 were originally positioned. While these new positions of F121 and F122 are conserved throughout the remainder of the simulation, approximately 29 ns into the dynamics phase of the simulation the dsRNA on the 3'OH side of the nick again moved away from the protein. These structural fluctuations indicate that the dsRNA undergoes occasional “breathing” motions, moving toward and away from the protein on the nanosecond timescale. During these breathing motions, F121 and F122 may move from positions where they intercalate the minor groove to positions where they make non-specific van der Waals interactions with the ribofuranose moieties flanking the minor groove. This suggestion is consistent with polynucleotide temperature (Debye-Waller) factors reported for the DNA/RNA hybrid in the T4 Rnl2 complex.¹⁶⁵

RNA binding through a unique kinetoplastid insert region. The kinetoplastids share a unique insert region that roughly spans *TbREL1* residues 165 to 200, depending on the exact alignment (Figure 14).^{86, 87, 165} Two regions of this insert lack defined secondary structure and contain positively charged residues that were postulated to be involved in RNA binding.⁸⁶ Based on a homology model to the T4 Rnl2 structure, Nandakumar et al. hypothesized that the loops may rearrange or adopt secondary structure and help stabilize the 3'OH side of the nicked dsRNA.¹⁶⁵ More recently, ensemble averaged electrostatic calculations over a 20 ns molecular dynamics simulation of the ATP-bound *TbREL1* protein without RNA indicated a large lobe of positive electrostatic potential near this region, further substantiating the suggested RNA-binding role.⁸⁷ The electrostatic calculations also suggested that charged residues K166, K172, K175, and/or R194 could play important dsRNA-stabilizing roles.⁸⁷

In the current work, we directly monitor the dynamics of this unique *TbREL1* insert region and assess its dsRNA-binding capacity. Through a set of transient interactions, the insert region significantly impacts the mobility of the gRNA strand, thus providing a secondary stabilization effect on the nicked 3'OH strand (Figures 18 and 19). Despite the prediction that K166, K172, and K175 would play important dsRNA stabilizing roles, the current simulation shows that none of these residues directly contact the RNA. However, because they are commensurate with a large lobe of positive potential,⁸⁷ they most likely participate in providing the electrostatic environment responsible for directing the diffusional association of the RNA substrate

into a productive bimolecular association complex. R194, on the other hand, directly contacts the RNA, forming a persistent salt-bridge with the non-bridging phosphate oxygen atoms of the gRNA strand backbone (Table 7). Despite their positive charge, K183 and K184, two residues positioned adjacent to the terminal end of the 3'OH strand, do not make persistent salt bridges or hydrogen bonds with the dsRNA. The RNA/DNA hybrid bound to T4 Rnl2, which was used to construct the simulated model, likewise does not take advantage of the high positive electrostatic potential immediately adjacent to this region; thus we note that the observed behavior may be an artifact of our homology model. Yet, the stability conferred by the kinetoplastid insert on the bound RNA is apparent (Figure 19). The kinetoplastid insert contacts the gRNA sequence from positions +2 to +4, corresponding to a 0.5 Å drop in RMSF, and contacts the nicked 5'PO₄ strand at positions +10 and +11, attenuating fluctuations at the 3'OH terminal.

6.4 Conclusions

The current molecular dynamics simulation of *Tb*REL1 in complex with a dsRNA substrate has clarified the roles of active-site residues during step-three of catalysis and allowed us to suggest a possible deprotonation pathway involving a non-bridging oxygen of the AppN intermediate. In an effort to help clarify the number of metals required for catalysis, a Mg²⁺ was positioned within the active site based on homology with other superfamily members. By examining the relation of the Mg²⁺ to the reactive residues, we inferred that Mg²⁺ is likely to confer transition-state

stabilization during nicked-strand religation and may also help stabilize the ribofuranose of the AppN intermediate. Both of these functional aspects have the potential to impart greater catalytic efficiency; as a result, we conclude that the *TbREL1* mechanism likely requires two Mg^{2+} for optimal efficiency. Additionally, by examining the persistence contact times between the protein and RNA during our simulation, we were able to explain, at least in part, why *TbREL1* requires RNA on the 5'PO₄ side of the nick *in vitro*¹⁷⁹ and to clarify the role of the unique kinetoplastid insert in RNA binding.

Our analysis allows us to make several predictions that may motivate future experimental work, adding to a growing body of knowledge concerning the structure-function relationship of RNA-binding residues within this enzyme. The side chains of D62, S65, S67, and Q70 make RNA-specific contacts; as a result, we predict that mutations to these residues will adversely affect the religation rate by disrupting RNA binding. Additionally, the recognition sequence formed by D62, L63, P64, S65, S67, and Q70 collectively form a smooth surface that intercalates the minor groove; consequently, we predict that mutating any of these residues to a large, bulky residue like tryptophan will perturb the protein-surface continuity and attenuate the religation rate by lowering RNA affinity. Positively charged K166, K172, and K175 reside within the unique kinetoplastid insert loop but do not directly contact the RNA. Instead, they are predicted to contribute to the electrostatic environment responsible for mediating RNA association. As a result, we predict that mutating K166, K172, and K175 to alanine will lower the religation activity by decreasing the total charge on

the protein. We note that while step-one autoadenylation requires ATP binding, the net charge on ATP is significantly less than that carried by dsRNA, and as a result step-one activity will be significantly less influenced by the decrease in total protein charge.

6.5 Acknowledgements

The National Center for Supercomputing Applications, the San Diego Supercomputing Center, the W.M. Keck Foundation, Accelrys, Inc., the National Biomedical Computational Resource and the Center for Theoretical Biological Physics are gratefully acknowledged.

Chapter 6, in full, is a near reprint of the material as it appears in *Biochemistry* 2009, volume 3, pages 709-719. The dissertation author and Robert V. Swift were the primary co-contributors to this work. Rommie E. Amaro also made significant contributions, and J. Andrew McCammon served as head adviser and principal investigator. The American Chemical Society holds the copyright for this manuscript and has authorized its publication in this dissertation.

CHAPTER 7: A Multidimensional Strategy to Detect Polypharmacological Targets in the Absence of Structural and Sequence Homology

Jacob D. Durrant, Rommie E. Amaro, Lei Xie, Michael D. Urbaniak, Michael A. J. Ferguson, Antti Haapalainen, Zhijun Chen, Anne Marie Di Guilmi, Frank Wunder, Philip E. Bourne, and J. Andrew McCammon

7.1 Introduction

Researchers have traditionally focused *in silico* efforts on designing inhibitors of specific protein targets, giving less attention to the computational identification of unpredicted secondary targets. This tendency is surprising given the frequency with which secondary receptors are responsible for both detrimental and beneficial pharmacological effects. The cost of developing a novel therapeutic ranges from \$500 million to \$2 billion dollars.¹⁸³ Millions of dollars are typically invested to advance a compound through clinical trials, but one third of these compounds fail or are later removed from the market due to unacceptable, medically harmful side effects³² often caused by binding to off-target receptors. The detrimental effects caused by drug binding to unknown secondary targets can be financially and clinically devastating.

In other cases, compound binding to multiple therapeutic targets (polypharmacology) is beneficial. Conventional drug discovery embraces the “one gene, one drug, one disease” philosophy; however, drugs that target only one protein are susceptible to resistance, as a single amino-acid mutation in the target active site often substantially reduces compound binding affinity. Resistance to multi-target drugs, on the other hand, requires simultaneous mutations in multiple protein targets. Furthermore, drugs with polypharmacological mechanisms often have greater

therapeutic potency. Some serotonergic drugs, for example, bind both 5-HT G-protein coupled receptors as well as the 5-HT_{3A} ion channel to achieve their therapeutic benefits, despite the fact that these two target proteins are not related by sequence or structure.¹⁸⁴

Identifying secondary targets in neglected tropical diseases, diseases for which drug development is neither profitable nor prevalent, allows researchers and doctors to retool approved drugs as novel treatments for the otherwise abandoned infections of the developing world. For instance, eflornithine, initially developed as an anti-cancer compound, was found to be a potent inhibitor of *Trypanosoma brucei* ornithine decarboxylase and is now a critical therapeutic in the fight against human African trypanosomiasis.¹⁸⁵ Examples like these motivate the need to develop new tools and algorithms to predict potential protein targets of candidate compounds.

A barrier to the development of these tools is the frequent absence of apparent evolutionary relationships among the multiple protein targets of a given compound, requiring that any potential method be capable of identifying target receptors independently of global sequence or structural homology. One approach is chemo-centric¹⁸⁶; similar ligands are more likely to have similar properties and therefore often bind proteins with similar active sites. A number of studies have successfully identified secondary receptors by comparing their known small-molecule ligands,^{184, 187-189} leading to probabilistic models that can in some cases successfully predict polypharmacology. Despite these successes, however, chemo-centric approaches have their limitations. Chemically similar small molecules do not always inhibit proteins

with similar active sites; indeed, even small changes in the chemical structures of some small-molecule inhibitors can drastically alter binding affinity¹⁸⁴ and the broad profile of binding to pharmacological targets.

A second approach is protein-centric. As the evolutionary relationships between secondary targets are not always apparent,¹⁸⁴ receptor active-site geometries and pharmacophores must be compared directly, independently of global sequence or structural homology. Geometric constraints have been used extensively to identify binding sites and to assess binding-site similarity.¹⁹⁰⁻¹⁹⁷ Of these methods, the sequence order independent profile-profile alignment (SOIPPA) algorithm is largely insensitive to both conformational changes in protein structure as well as the uncertainty inherent in homology models and low-resolution structures.¹⁹⁸

In the current work, we present a novel multidimensional strategy to identify the multiple protein receptors of a given compound that incorporates three levels of information: sequence-based homology clustering, the SOIPPA algorithm in conjunction with a geometric-potential metric,¹⁹⁸⁻²⁰⁰ and *in silico* ligand docking. To demonstrate the utility of the strategy, we identify several human and pathogen secondary targets of compound **1** (NSC-45208), 4,5-dihydroxy-3-(1-naphthyldiazenyl)-2,7-naphthalenedisulfonic acid, a recently discovered micromolar inhibitor of *T. brucei* RNA editing ligase 1 (*TbREL1*).³¹ *TbREL1* is a confirmed drug target in *T. brucei*, the causative agent of human African trypanosomiasis, a disease for which drug development has been largely neglected.²

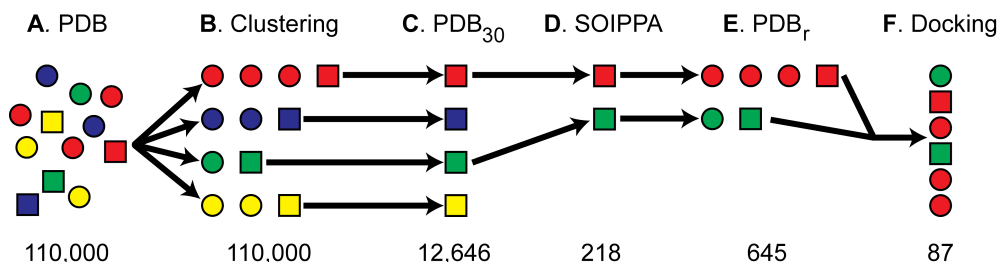


Figure 20: The strategy workflow. Circles and squares represent protein chains. Homologous chains share the same color. From each group of homologous chains, one is selected as representative and is shown as a square. (A) As the PDB has approximately 110,000 protein chains, identifying secondary targets directly is computationally intractable. (B) To reduce the number of chains, all chains are grouped by sequence homology into 12,646 clusters, and (C) a single representative chain is selected from each cluster. The set of all representative chains is called the PDB₃₀. (D) SOIPPA is used to eliminate all protein chains in the PDB₃₀ with active sites that are dissimilar to that of the primary target, *TbREL1*. Only 218 chains remain. (E) A new set of 645 protein-chain structures called the PDB_r is created by taking the union of all those clusters whose representative PDB₃₀ protein chains have active sites that are not dissimilar to that of *TbREL1*. (F) Redundant chains are ignored; compound 1 is docked into the remaining 87 protein chains. Chains are ranked by their corresponding docking scores.

7.2 Results

Predicted Secondary Targets

Sequence homology clustering was used to identify 12,646 protein-chain clusters from among the 110,000 protein chains present in the Protein Data Bank (PDB) as of late 2007 (Figure 20A, B). A representative protein chain was chosen from each cluster, creating a smaller set of chains called the PDB₃₀ (Figure 20C). The SOIPPA algorithm in conjunction with a geometric-potential metric¹⁹⁸⁻²⁰⁰ revealed that the active sites of 12,428 of the representative PDB₃₀ protein chains (98.3%) were dissimilar to that of *Tb*REL1 ($p > 0.05$), the known target. After discarding these dissimilar protein chains, 218 representative PDB₃₀ chains remained (Figure 20D). The remaining PDB₃₀ proteins and the clusters they represented were merged into a single list containing 2,897 chains, a list enriched with possible secondary targets (Figure 20E). By considering only proteins from human or known human-pathogen species and eliminating PDBs with formatting errors, the number of chains was reduced from 2,897 to 645 (22.3%). This new list of protein chains was called the PDB_r (PDB_{reduced}).

Compound **1**, a known inhibitor of the primary target, *Tb*REL1, was docked into each of the 645 potential secondary targets of the PDB_r with AutoDock 4.0 (Figure 20F). After docking, both protein chains of unknown function and chains belonging to proteins with duplicate names were deleted without regard for species. This pruning produced a list of 87 non-redundant predicted secondary targets. The docked pose of **1** into each of these 87 predicted secondary targets was analyzed to

confirm binding to a pocket of known biochemical or pharmacological activity. In two instances, SITE data was included in the published PDB file, allowing us to identify one active-site and one off-site docking. In 25 instances, the docked ligand bound at the same location as a co-crystallized ligand, typically superimposed on top of it. Homology modeling revealed that an additional nine docked ligands bound at active sites of known biochemical or pharmacological activity, while 34 bound at alternate sites on the receptor. Though these alternate sites may be involved in allosteric regulation of protein function, we chose not to pursue them further. In all, the list of 87 protein chains contained 35 chains whose known active sites contained docked ligands, 35 chains whose alternate sites contained docked ligands, and 17 chains that could not be classified.

Experimental Validation of Predicted Secondary Targets

The predicted secondary targets that gave the best docking scores, *H. sapiens* mitochondrial 2-enoyl thioester reductase (*HsETR1*), *T. brucei* UDP-galactose 4' epimerase (*TbGalE*), *H. sapiens* phosphodiesterase 9A (*HsPDE9A2*), and *Streptococcus pneumoniae* teichoic acid phosphorylcholine esterase (*SpPce*), were subsequently tested experimentally.

Table 8: Selected predicted secondary targets of compound 1 in humans. Human secondary targets are involved in metabolism; polynucleotide synthesis, repair, and replication; amino acid synthesis; blood clotting; vision; and nuclear transport. “AD score” refers to the AutoDock-predicted energy of binding to 1; “SOIPPA *p*-value” refers to the similarity between the secondary-target and *TbREL1* active sites; “sequence identity” refers to the percent amino-acid identity with *TbREL1* as computed by the CLUSTALW2 algorithm; and “FATCAT *p*-value” refers to the structural similarity between the secondary target and *TbREL1*. Protein sequences were extracted from PDB *seqres* headers.

	Protein	AD Score	SOIPPA <i>p</i> -value	Sequence Identity	FATCAT <i>p</i> -value
Metabolism	UDP-glucose 4-epimerase (1I3L:B)	-11.22		3%	0.613
	NAGK (2CH5:C)	-10.75	3.45×10^{-2}	9%	0.351
	acetyl-CoA carboxylase 2 (2HJW:A)	-10.28	7.82×10^{-3}	5%	0.423
	mitochondrial 2-enoyl thioester reductase (1ZSY:A)	-10.04	$< 1 \times 10^{-5}$	4%	0.754
	tubby isoform A (1S31:A)	-9.17		6%	0.615
DNA synthesis, repair, replication	DNA ligase I (1X9N:A, residues 535-751)	-9.70	$< 1 \times 10^{-5}$	5%	3.82×10^{-3}
	3-methyl-adenine DNA glycosylase (1EWN:A)	-9.09	1.18×10^{-2}	10%	0.800
	thymidylate synthase (1I00:A)	-8.50		3%	0.237
Amino acid synthesis	pyrroline-5-carboxylate reductase 1 (2GR9:B)	-10.49	4.28×10^{-2}	1%	0.702
Blood clotting	fibrinogen (1FZE:B)	-9.53		4%	0.626
Vision	tubby related 1 (2FIM:B)	-10.21	2.60×10^{-2}	3%	0.637
Nuclear transport	snurportin-1 (1XK5:A)	-10.10	$< 1 \times 10^{-5}$	4%	6.07×10^{-4}

Compound **1** inhibited *HsETR1* with a measured IC_{50} of approximately 33.5 μ M and a Hill slope of 1.06. Neither the FATCAT algorithm nor CLUSTALW2 judged *HsETR1* to be significantly homologous in sequence or structure to the primary target, *TbREL1* ($p = 0.754$; identity = 4%). In contrast, the SOIPPA algorithm judged the active sites of these two proteins to be significantly similar ($p < 1 \times 10^{-5}$) (Table 8). *HsETR1* aggregates were not detected, as measured by both spectrometry and centrifugation. A mixture of compound **1** and *HsETR1* was run through a gel

filtration column, thereby removing any unbound ligand. Spectroscopic analysis subsequently demonstrated that **1** was not covalently bound to *HsETR1*.

Table 9: Selected predicted secondary targets of compound **1** in pathogens. “AD score” refers to the AutoDock-predicted energy of binding to **1**; “SOIPPA *p*-value” refers to the similarity between the secondary-target and *TbREL1* active sites; “sequence identity” refers to the percent amino-acid identity with *TbREL1* as computed by the CLUSTALW2 algorithm; and “FATCAT *p*-value” refers to the structural similarity between the secondary target and *TbREL1*. Protein sequences were extracted from PDB *seqres* headers.

Protein	Species	AD Score	SOIPPA <i>p</i> -value	Sequence Identity	FATCAT <i>p</i> -value
probable ATP-dependent DNA ligase (2FAO:A)	<i>Pseudomonas aeruginosa</i>	-10.57	1.34 X 10 ⁻²	5%	0.440
UDP-galactose 4-epimerase (1GY8:D)	<i>T. brucei</i>	-10.29		1%	0.622
dTDP-D-glucose 4,6-dehydratase (1KET:B)	<i>Streptococcus suis</i>	-9.55		5%	0.502
dihydrofolate reductase-thymidylate (1J3I:C)	<i>Plasmodium falciparum</i>	-9.53	1.34 x 10 ⁻²	5%	0.428
DNA ligase, NAD-dependent (1TAE:B)	<i>E. faecalis</i> v583	-9.49		8%	3.56 x 10 ⁻²
dTDP-D-glucose 4,6-dehydratase (1G1A:C)	<i>S. enterica</i>	-9.24		1%	0.529
adenine phosphoribosyltransferase (1MZV:A)	<i>Leishmania tarentolae</i>	-8.61		11%	0.649
UTP-gluc-1-P uridylyltransferase 2 (2OEG:A)	<i>Leishmania major</i>	-8.56	7.82 x 10 ⁻³	8%	0.724
purine nucleoside phosphorylase (2B94:A)	<i>Plasmodium knowlesi</i>	-7.61		2%	0.650
DNA ligase (1ZAU:A)	<i>M. tuberculosis</i>	-6.75		4%	2.82 x 10 ⁻²

Compound **1** inhibited *TbGale* with a measured IC₅₀ of 0.7 ± 0.2 μM and a Hill slope of 1.13 +/- 0.36. Again, the FATCAT algorithm did not judge the structure of *TbGale* to be significantly similar to that of the primary target, *TbREL1* (*p* = 0.627), and the CLUSTALW2 algorithm did not identify significant sequence homology (identity = 1%) (Table 9). *TbGale* inhibition was unaffected by the

presence of detergent, and activity could be restored by dialysis of the protein, demonstrating that the *Tb*GaIE inhibition was not due to aggregation of the compound or chemical modification of the protein.

Two of the predicted secondary targets, *Hs*PDE9A2 and *Sp*Pce, were uninhibited by **1** at 200 μ M and 10 mM, respectively. AutoDock predicted that **1** would bind *Hs*PDE9A2 and *Sp*Pce with -18.19 and -28.00 kcal/mol, respectively (Tables 1 and 2).

7.3 Discussion

In this work, we attempt to further the study of polypharmacological and side-effect prediction by presenting a multidimensional strategy for identifying secondary targets of known enzyme inhibitors in the absence of global structure and sequence homology. To demonstrate the utility of the strategy, we identify secondary targets of **1**, a recently discovered inhibitor of *Tb*REL1³¹ from *T. brucei*, the causative agent of African sleeping sickness. *Tb*REL1 plays a critical role in the editing of trypanosomatid mitochondrial RNA transcripts and is required for the survival of both the *T. brucei* insect and bloodstream forms.^{89, 90} Additionally, *Tb*REL1 is a particularly attractive drug target because there are no known close human homologs.³¹

Compound **1** was chosen to illustrate how the current strategy can be applied early in the drug-discovery process. The compound inhibits a known drug target, satisfies Lipinski's rule of five, and is structurally similar to surinam, a drug currently approved for the treatment of human African trypanosomiasis. In this sense, **1** is drug

like. Compound **1** has not yet been optimized to bind *TbREL1* in the nanomolar regime, however, and does contain some undesirable functional groups, and so is still very much under development. By incorporating the identification of secondary targets early in the drug-design process, we hope to eventually make modifications to compound **1** that will increase the binding affinity to the primary target while decreasing binding to undesirable secondary targets.

Human Secondary Targets

Of the 35 predicted secondary targets of compound **1** identified, twelve were human proteins. Potential side effects of **1** can be predicted by considering the physiological role of these targets. For instance, a number of the predicted secondary human targets regulate metabolism, including the experimentally confirmed secondary target *HsETR1*. Neither FATCAT nor CLUSTALW2 judged *HsETR1* to be homologous to the primary target, *TbREL1* (Table 8). The current strategy, which is not dependent on sequence or global structural homology, was able to identify this secondary target where identification by homology would have failed. *HsETR1* is thought to be essential for fatty acid synthesis (FAS) type II.²⁰¹ In the process of optimizing **1** to make it more druglike, modifications that reduce binding to human *HsETR1* may diminish unforeseen side effects. Interestingly, AutoDock predicted that **1** partially occupies a co-factor (NADPH) binding site, suggesting that the compound may function as a competitive inhibitor for NADPH.

H. sapiens UDP-galactose 4-epimerase (*HsGalE*), a second protein involved in human metabolism, was also identified as a potential secondary target. Though *HsGalE* shares little homology with the primary target, *TbREL1* (FATCAT *p*-value: 0.610; CLUSTALW2 identity: 2%), it is highly homologous with *TbGalE* (FATPAT *p*-value: 0.00; CLUSTALW2 identity: 37%), which we show to be a secondary target of **1** (IC₅₀: 0.7 ± 0.2 μ M). Mutations in the *HsGalE* gene cause type 3 galactosemia in humans. As toxic levels of galactose build up in patients' blood, vomiting, hepatomegaly, jaundice, renal failure, and cataracts typically follow.²⁰² Chronic administration of **1** is thus ill advised, though short-term treatment may be acceptable if patients avoid dairy and other sources of galactose.

In addition to metabolism, a number of the predicted human secondary targets are involved in polynucleotide synthesis, repair, and replication. *H. sapiens* DNA ligase I (*HsLigI*), a protein that belongs to the same enzyme superfamily as *TbREL1*, is one notable example. *HsLigI* catalyzes the ultimate, essential step in DNA replication, repair, and recombination.²⁰³ Studies have demonstrated that *HsLigI* is defective in at least one representative lymphoid cell line of Bloom's syndrome origin,²⁰⁴ and a mouse model with a mutant *HsLigI* allele exhibits an increased incidence of spontaneous cancer.²⁰⁵

Table 10: Global sequence and structural homology between *HsLigI*, residues 535 to 751, and three other selected DNA ligases.

	Identity Similarity	FATCAT <i>p</i> value
<i>HsLigIIIβ</i>	31%	-
<i>E. faecalis</i> v583 NAD-dependent DNA ligase	4%	2.36×10^{-7}
<i>M. tuberculosis</i> DNA ligase	4%	1.07×10^{-6}

Though not tested explicitly, additional evidence does suggest that **1** binds *HsLigI*. First, **1** is known to inhibit *H. sapiens* DNA ligase IIIβ (*HsLigIIIβ*; IC₅₀: 27.49 ± 6.40 μM),³¹ a *HsLigI* homolog (Table 10). Second, the FATCAT algorithm²⁰⁶ judged the structure of *HsLigI* to be significantly similar to that of the primary target, *TbREL1*, and the SOIPPA algorithm judged the active sites of these two proteins to be similar (Table 8). Finally, the AutoDock-predicted binding energy of **1** to *HsLigI* was -9.70 kcal/mol (Table 8).

Secondary Targets in Bacterial and Parasitic Pathogens

Of the 35 predicted secondary targets identified, 23 belonged to bacterial and parasitic species. Among these predicted pathogenic secondary targets, enzymatic assays demonstrated that **1** inhibits *TbGale*⁴⁵ (Table 9) with high nanomolar affinity. Interestingly, the FATCAT algorithm did not judge the structure of *TbGale* to be significantly similar to that of the primary target, *TbREL1* (Table 9), nor did CLUSTALW2 suggest sequence homology (identity = 1%). Thus, had we attempted to identify secondary targets by protein sequence or structural homology alone,

TbGalE would not have been detected. This result again illustrates the power of the current strategy.

Like *TbREL1*, *TbGalE* is essential for *T. brucei* survival.⁴⁷ Thus, **1** inhibits two essential *T. brucei* enzymes, an example of potential polypharmacology. In the process of optimizing **1** to make it more druglike, modifications that increase binding to both *TbREL1* and *TbGalE* would likely improve drug efficacy and decrease the chances of resistance through mutation. Interestingly, AutoDock again predicted that **1** would bind partially in the *TbGalE* NAD⁺ pocket, suggesting that the compound may be a competitive inhibitor for NAD⁺ rather than UDP-galactose.

Though not tested explicitly, another predicted secondary target in pathogens, *Salmonella enterica* dTDP-D-glucose 4,6 dehydratase (*SeRmlB*), shares sequence and structural homology with *TbGalE* and may also bind **1** (Table 9). The FATCAT algorithm judged *SeRmlB* to be significantly structurally similar to *TbGalE* (*p* value = 0.00), and CLUSTALW2 identified some sequence homology (identity = 21%). Additionally, AutoDock predicted that *SeRmlB* would bind **1** with relatively high affinity (Table 9).

SeRmlB is the second enzyme in the dTDP-L-rhamnose biosynthetic pathway; L-rhamnose is part of the LPS endotoxin (the O antigen) in many serotypes and serovars of Gram-negative bacteria. As L-rhamnose is common in the cell walls and envelopes of some pathogenic bacteria but absent in humans, *SeRmlB* is thought to be a potential drug target.²⁰⁷ These findings support the hypothesis that compounds similar to **1** may have anti-bacterial properties.

Aside from inhibiting an essential protein in the bacterial dTDP-L-rhamnose biosynthetic pathway, the current strategy also identified two bacterial DNA ligases as potential secondary targets. *Enterococcus faecalis* v583 NAD-dependent DNA ligase (*Ef* ligase) and *Mycobacterium tuberculosis* DNA ligase (*Mt* ligase) are structurally homologous to *HsLigI* (Table 10). Human DNA ligases require ATP as a co-factor, but bacterial ligases require NAD^+ .²⁰⁸ Because of this important biochemical difference, bacterial ligases are thought to be good drug targets.

Method Limitations

A limitation of the current method is the prediction of false positives. Two of the predicted secondary targets, *SpPce* and *HsPDE9A2*, were uninhibited by **1**. Closer inspection of the docking results revealed electrostatic energy components of -30.89 kcal/mol and -12.29 kcal/mol, respectively. In both cases, the partial charges of several active-site metal cations had been manually set to the corresponding formal charge. Subsequent analysis of the docked poses revealed that for both receptors, one of the sulfonate groups of **1** was juxtaposed against these highly charged metal cations. Clearly, a more careful treatment of the electrostatic interactions that accounts for electron polarization is warranted when docking into active sites that include metal ions.

An additional limitation of the current method is its dependence on sequence-homology clustering. Because of computational limitations, the number of potential secondary targets analyzed with SOIPPA had to be reduced; consequently, rather than

analyzing all protein chains in the PDB, representative chains were selected from clusters of homologous proteins based on the supposition that sequence-homologous proteins would likely have similar active sites. This supposition, however, is hardly a universal truth. Through convergent evolution, two proteins with very different primary sequences may have evolved to bind similar ligands, and so may have similar active sites despite a lack of sequence homology.

Fortunately, recent advances in the SOIPPA algorithm now make sequence-homology clustering unnecessary. The version of SOIPPA used in the current study estimated the statistical significance of each active-site comparison using a non-parametric statistics method that required at least several hundred additional comparisons to derive a background distribution. Recently, an extreme-value distribution model has been developed to compute the statistical significance of SOIPPA scores²⁰⁹; this model improves the speed of the algorithm by at least two orders of multitude, so that each active-site comparison can be performed in mere seconds. Ligand binding-site similarity searches can now be performed on a genome-wide scale without the need for sequence-homology clustering.^{209, 210} The new statistical model has been implemented in SMAP v2.0, available at <http://funsite.sdsc.edu>.

The manual verification of predicted compound binding to active sites of known biochemical or pharmacological activity also presented a limiting challenge. This step was very time consuming; had multiple compounds been tested, manual verification of all docked poses may have been impossible. To automate the process,

active-site binding can be confirmed in many cases based on the proximity of the docked ligand to known catalytic residues annotated in the Catalytic Site Atlas²¹¹ or to active-site residues specified in the site records of the PDB. Additionally, the SOIPPA implementation in SMAP v2.0, not available until recently, suggests an initial ligand binding pose for each predicted secondary target. Docked poses could be compared to this initial suggestion using an automated script and rejected in the absence of proximity.

Accounting for Possible Promiscuous Binding

As experimental validation has confirmed that **1** inhibits multiple protein targets, the possibility of nonspecific, promiscuous binding must be eliminated. Previous work has demonstrated that compound **1** inhibition of several ATP-dependent proteins depends on the degree of homology with the primary target, *TbREL1*, suggesting that binding occurs *via* a specific rather than promiscuous mechanism.³¹ Additionally, the lack of *SpPce* or *HsPDE9A2* inhibition further suggests that indiscriminant inhibition is unlikely.

Promiscuous inhibition can occur when a compound forms colloidal aggregates that inhibit indiscriminately. In one recent study, 95% of the inhibitors identified in a high-throughput screen were subsequently found to inhibit *via* a nonspecific aggregation-based mechanism.⁶⁸ In theory, the chances of aggregation are reduced in the case of **1** because of its negative charge; individual molecules should repel each other, preventing aggregation. To confirm this theory, Amaro et. al assayed

1 against *TbREL1* in the presence of a non-ionic detergent as well as an additional protein (BSA) known to prevent aggregation. The presence of the detergent and the separate test with BSA did not significantly influence *TbREL1* inhibition, demonstrating that **1** does not aggregate.³¹

The current work likewise suggests no aggregation. The presence of detergent had no effect on *TbGalE* inhibition. Additionally, no *HsETR1* aggregation was observed, as measured by both spectrometry and centrifugation. The Hill slopes corresponding to the inhibition of *HsETR1* and *TbGalE* were 1.06 and 1.13, respectively. As these slopes are approximately equal to one, the inhibition of these two proteins likely occurs *via* ligand binding to a single site, as predicted. One recent study suggested that aggregation-based inhibition typically produces Hill slopes that are much steeper, with average values of about 2.2.⁶⁸

Promiscuous inhibition can also occur when compounds chemically modify the proteins they inhibit. Several experiments were performed in order to rule out chemical modification. To test for chemical modification of *TbGalE*, *TbGalE* was incubated with **1**, and subsequent dialysis was used to remove any unbound ligand. *TbGalE* activity was unaffected following dialysis, as compared to treatment with DMSO alone. Had **1** been covalently linked to *TbGalE*, the compound would not have been washed away, and the enzyme would have shown little activity.

To test for chemical modification of *HsETR1*, a mixture of compound **1** and *HsETR1* was run through a gel filtration column, which likewise removed any unbound ligand. The absorption spectrum of the fraction containing *HsETR1* was

subsequently analyzed and did not demonstrate the peaks characteristic of **1** at 530 nm and 320 nm, likewise demonstrating that **1** was not covalently bound to *HsETR1*.

A cursory Chemo-Centric Analysis

A number of the predicted secondary targets identified belong to the same or similar biochemical pathways (i.e. metabolic pathways; DNA synthesis, repair, and replication pathways; and DNA ligase pathways) (Table 8). This result is encouraging, as proteins of the same pathway often act on similar substrates and so have similar active sites. A cursory chemo-centric look at the native substrates of the identified secondary targets in many instances corroborates our findings. For instance, the human proteins *HsLigI*, 3-methyl-adenine DNA glycosylase, and thymidylate synthase, all involved in DNA synthesis, repair, and replication, were identified as potential target receptors. *HsLigI* is a *TbREL1* homolog (Table 8) that ligates DNA in a way analogous to *TbREL1* dsRNA ligation. In contrast, 3-methyl-adenine DNA glycosylase is not a *TbREL1* homolog (Table 8), but an examination of its structure nevertheless reveals that it also binds DNA. One of the nucleotides of the bound DNA, an alkylated base generated endogenously by lipid peroxidation, protrudes into a deep binding pocket in a way analogous to *TbREL1*-ATP binding.²¹² Thymidylate synthase, another predicted secondary target involved in nucleotide synthesis, does not bind double-stranded nucleic acid, but rather binds deoxyuridine monophosphate, a compound with a nucleotide-ribose-phosphate substructure like that of ATP, a known substrate of the primary target, *TbREL1*.

Notably, two experimentally validated secondary targets contain NAD⁺ or NADPH co-factors, and in both cases **1** is predicted to bind at least partially in one of the co-factor binding pockets. Similar to ATP, NAD⁺ and NADPH both contain adenine-ribose-biphosphate substructures. Of the 23 proteins listed in Tables 1 and 2, eight contain NAD⁺-like co-factors.

Traditionally, researchers have devoted their computational efforts to designing inhibitors of specific protein targets while paying less attention to the *in silico* prediction of secondary targets. Because adverse side effects are often discovered late in the drug-development process, often after the investment of many millions of dollars, we recommend using the current strategy to help bad drugs “fail early,” or, better yet, to guide the drug-discovery process towards more selective inhibitors. Additionally, the current methodology can help medicinal chemists overcome the conventional “one gene, one drug, one disease” paradigm. The rational design of drugs that act *via* polypharmacological mechanisms can produce compounds that exhibit increased therapeutic potency and against which resistance is less likely to develop.

7.4 Materials and Methods

Compound Purity

Compound **1** was provided by the NCI/DTP Open Chemical Repository (<http://dtp.cancer.gov>). Compound identity was confirmed by high-resolution mass spectrometry, and no impurities were detected by ¹H-NMR.

Computational Methodology

The computational strategy presented here utilizes three distinct components in order to identify secondary pharmacological targets.

1. Homology Clustering: Creating a Non-Redundant PDB Representation

Identifying potential secondary targets from among the 110,000 protein chains deposited in the RCSB PDB¹⁴² as of late 2007 was judged computationally intractable. In order to reduce the number of protein chains, redundancies in the RCSB PDB were eliminated by clustering all protein chains by sequence using the NCBI *blastclust* program, with a sequence identity threshold of 30% and an overlap threshold of 0.9 (Figure 20A, B). A representative protein chain was then chosen at random from each cluster, thus creating a smaller set of chains called the PDB₃₀ (Figure 20C).

2. SOIPPA: Identifying Active Sites Similar to that of the Primary Target

In order to eliminate those members of the PDB₃₀ whose active-site alpha-carbon configurations were different enough from that of the primary target, *TbREL1*, so as to likely preclude NSC4520 binding, we used the SOIPPA algorithm in conjunction with a geometric potential,¹⁹⁸⁻²⁰⁰ which computes general binding-site similarity based on shape (alpha-carbon tessellation), physical properties, and the evolutionary profiles of the active-site residues, without regard for specific side-chain positions or global sequence or structure (Figure 20D). Using SOIPPA and the

geometric potential, we eliminated all protein chains in the PDB₃₀ with active sites that were dissimilar to that of the primary target, *TbREL1* (p -value > 0.05) (Figure 20D). The p -value was calculated from a non-parametric density function generated from 980 PDB chains with unique SCOP folds.¹⁹⁹ To derive the background distribution, a Gaussian function was placed at each observation. The mean of the Gaussian of the observed binding-site similarity score and its variance were fixed. The final probability density function was the sum of all these Gaussian functions. The optimal bandwidth was estimated from the data by using a least square cross-validation approach.²¹³

Each representative protein chain corresponded to a PDB cluster containing multiple homologous chains. A new set of protein-chain structures called the PDB_r (PDB_{reduced}) was created by taking the union of all those clusters whose representative PDB₃₀ protein chains had active sites that were not dissimilar to that of *TbREL1* (p -value < 0.05) (Figure 20E). By considering only proteins from human or known human-pathogen species, the number of chains was significantly reduced. An additional protein, 1GJJ, was eliminated because of apparent PDB formatting errors. 1S31 was retained despite having a malformed GLU residue (561) to which Gasteiger partial charges could not be assigned.

3. In silico *Docking*

AutoDock 4.0²⁶ was used to dock **1** into the protein chains of the PDB_r (Figure 20F). In previous work, AutoDock was validated against *TbREL1*.³¹ To define the distinct docking grid associated with each protein chain, SOIPPA was used to identify the most probable active site, and the grid box was set to the smallest possible X-Y-Z cube encompassing all the atoms of the SOIPPA-reported alignable alpha carbons.

All dockings were sorted by the predicted binding energy of their most-populated AutoDock clusters. After eliminating protein chains of unknown function from the list, the data was grouped according to species and protein, revealing significant redundancy in the PDB_r. We selected the ligand-protein pair from each group with the best AutoDock score, producing a list of non-redundant ligand-protein pairs, one corresponding to each protein/species group.

For each of these ligand-protein pairs, we used one of several methods to determine if AutoDock placed **1** in an active site of known biochemical or pharmacological activity. First, SITE data included in the published PDB file identified several active sites. Second, co-crystallized ligands bound in native active sites were examined. Finally, homology modeling was used to determine the locations of active sites for the remaining protein chains, when possible. A protein chain was considered to be a “hit” if **1** had a high predicted energy of binding and if **1** was predicted to bind in an identified active site of known biochemical or pharmacological activity.

7.5 Acknowledgements

Resources provided by the National Center for Supercomputing Applications, the San Diego Supercomputer Center, the W.M. Keck Foundation, the National Biomedical Computation Resource, and the Center for Theoretical Biological Physics are gratefully acknowledged. Compound **1** was provided by the NCI/DTP Open Chemical Repository (<http://dtp.cancer.gov>). We also would also like to thank Professor Rik K. Wierenga and Professor J. Kalervo Hiltunen.

Chapter 7, in full, is a near reprint of the material as it appears in PLoS Computational Biology 2010, volume 6: e1000648. The dissertation author was the primary contributor to this work. Rommie E. Amaro, Lei Xie, Michael D. Urbaniak, Michael A. J. Ferguson, Antti Haapalainen, Zhijun Chen, Anne Marie Di Guilmi, Frank Wunder, and Philip E. Bourne also made significant contributions, and J. Andrew McCammon served as head adviser and principal investigator. PLoS journals have an open-access policy that permits republication of PLoS articles in academic theses.

CHAPTER 8: AutoGrow: A Novel Algorithm for Protein Inhibitor Design

Jacob Durrant, Rommie E. Amaro, and J. Andrew McCammon

8.1 Introduction

Due in part to the increasing availability of crystallographic protein structures as well as rapid improvements in computing power, the past few decades have seen an explosion in the field of computer-based drug discovery. Algorithms developed to identify, generate, and optimize potential ligands *in silico* by optimizing the ligand-receptor hydrogen bond, electrostatic, and hydrophobic interactions have contributed to the development of a number of FDA-approved drugs.⁴⁴

Many computer-aided drug design techniques have been developed in recent years (reviewed in reference 214). Among these, comprehensive physics-based approaches, such as thermodynamic integration,⁹¹ single-step perturbation,⁹² and free energy of perturbation,⁹³ can yield very accurate binding free energies but are limited in the diversity of compounds generated and come at a high computational expense. Even so, physics-based ligand optimization methods based on density functional theory²¹⁵ and electron-nuclear attraction potentials²¹⁶ are promising new approaches.

Fragment-based growing strategies, which create novel structures by adding interacting moieties to a fixed scaffold, are generally more popular due to the lower computational expense and the higher degree of compound diversity generated.^{117, 217-225} As these algorithms typically limit moiety addition to a library of fragments, finding good ligands is computationally efficient. Moreover, the scoring functions used to predict the overall binding free energies of new ligands are generally faster,

though not as accurate as the comprehensive physics-based approaches described above. Notably, although many fragment-based growing strategies assume that the position or binding mode of the “core” scaffold does not change upon fragment addition, this assumption is not valid in many cases.

Ligand docking is another common drug development technique. Docking programs are commonly employed in virtual screening applications to search through databases of pre-defined compounds, position the ligands correctly within the receptor active site, and identify those ligands predicted to bind with high affinity. Developing and improving docking scoring functions is an area of active interest (see, for example, references 26, 95-97, 226-243). Like fragment-based growing scoring functions, docking scoring functions are typically faster and less accurate than more rigorous physics-based approaches. Unlike many growing techniques, docking typically accounts for the mobility of all portions of the ligand. However, rather than generate novel compounds, docking techniques are only able to suggest good binders from among those compounds contained in a database of pre-defined ligand candidates.

The current work is motivated by the desire to overcome the inherent weaknesses of both fragment-growing and docking algorithms. In this work, we introduce AutoGrow, a novel drug design algorithm that combines elements of both techniques. AutoGrow uses a growing strategy to build upon an initial “core” scaffold; molecular fragments are added at random to this scaffold, thereby generating a population of novel compounds. Rather than assuming that the common “core”

scaffold of the members of this population is static within the active site, AutoGrow dynamically redocks each novel compound into the protein receptor upon addition of varied fragments, producing *ab initio* poses for every molecule. An evolutionary algorithm then evaluates the docking scores of each population member, and the best binders become founders of the subsequent generation.

As generation after generation is created, each based on the most fit individuals of the previous generation, a larger inhibitor with a higher predicted binding affinity eventually evolves. While not necessarily drug-like, these predicted inhibitors often provide information that is useful early in the lead optimization process, including the identification of potential new protein-ligand interactions as well as novel drug scaffolds. To validate the AutoGrow algorithm, we recreate ATP, the known substrate of *Trypanosoma brucei* RNA editing ligase 1⁸⁶; oseltamivir, a nanomolar-affinity antiviral inhibitor of the neuraminidase enzyme¹¹¹; and aminoimidazole 4-carboxamide ribonucleotide (AMZ), a nanomolar-affinity inhibitor of AICAR transformylase.²²⁶

8.2 Methods and Materials

Evolutionary algorithms are ideally suited to complex problems such as those associated with *de novo* drug discovery.²²⁷ These algorithms typically include three operators, modeled on the three natural operators of biological evolution: selection, crossover, and mutation. The evolutionary procedure is divided into generations, where each generation consists of a population of individuals derived from selection

of the most fit members of the previous generation. The *internal* variation of each generation is exploited via crossover, wherein the characteristics of two “parent” individuals are combined to create a new “child” individual. *External* variation is introduced into each generation via mutation, wherein new individuals are created by making small, usually random changes to individuals already present in the population. As generation after generation is created, each based on the most fit individuals of the previous generation as well as additional individuals derived by exploiting internal and external variation, a solution eventually evolves. In the AutoGrow implementation, each generation consists of multiple potential ligands. The algorithm evolves ligands that are predicted to bind to a given target protein with high affinity.

AutoGrow uses AutoDock as the selection operator.²⁶ For each generation, all ligand files are docked into the target protein, and for each dock, AutoDock returns a predicted binding affinity. Those ligands that bind within the active site and have the most favorable predicted binding affinities are selected for inclusion in the next generation. AutoDock is a good selection operator; the program is free, takes into account full ligand flexibility, and has a well-tested scoring function.²²⁸ For these reasons, AutoGrow has been coded to interface with AutoDock by default. With only slight modification to the code, however, AutoGrow could use any number of docking programs as its selection operator. AutoGrow is open source, and we encourage source-code modifications of this type.

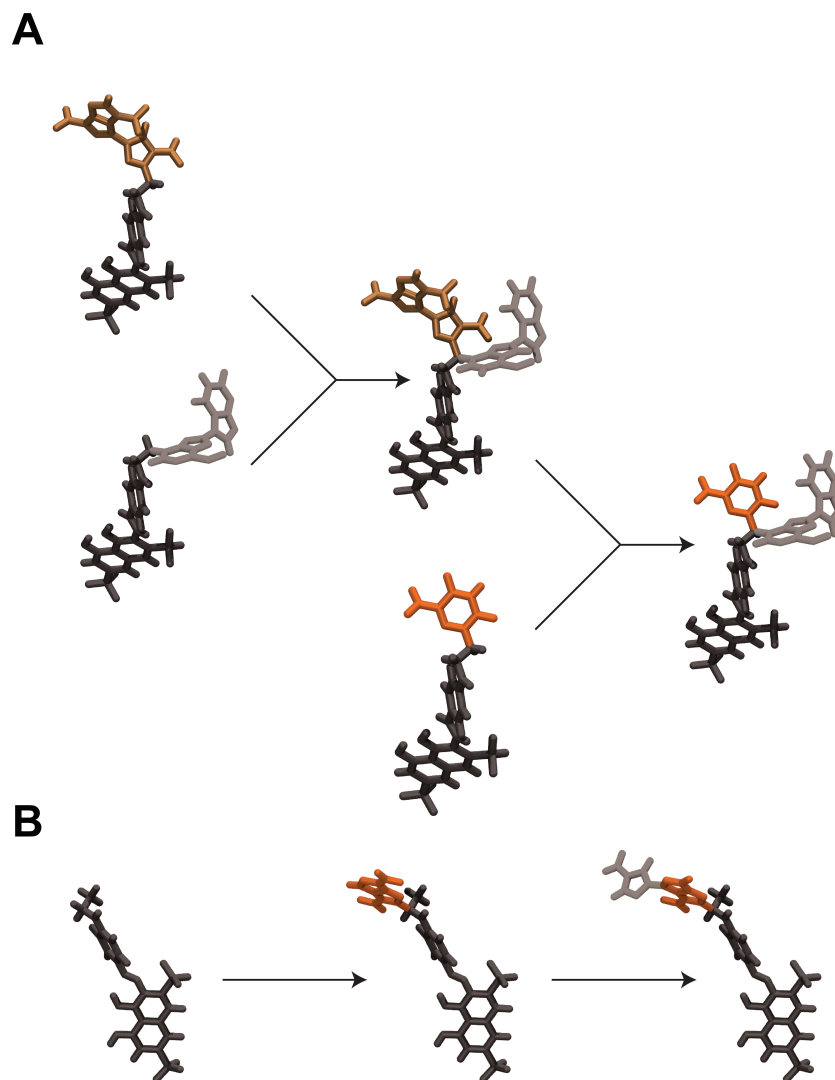


Figure 21: An illustration of the mutation and crossover operators. A) A crossover, or “child” ligand is formed by selecting an individual pair (“couple”) and creating a new hybrid ligand by randomly mixing and matching the moieties of the two “parents.” B) A “mutant” ligand is formed by replacing an appropriate scaffold linker hydrogen atom with a randomly chosen molecular fragment. Compound fragments are color coded to better illustrate crossover and mutation.

In order to exploit the *internal* variation present in each generation after the first, a crossover operator is used. First, individual pairs (“couples”) are selected from among those ligands that were the most fit of the previous generation. New hybrid “children” ligands are then formed by randomly mixing and matching the attached moieties of the two “parents” (Figure 21A).

In order to introduce *external* variation into each generation, a mutation operator is used. First, a number of individuals are selected from among those that were the most fit of the previous generation. For each of these individuals, an appropriate hydrogen atom (the “scaffold linker hydrogen”) is randomly selected. A molecular fragment is then chosen randomly from a fragment library, and one of the hydrogen atoms of that fragment is likewise selected at random (the “fragment linker hydrogen”). A new “mutant” ligand is created by linking the scaffold and fragment through their respective linker hydrogen atoms, replacing those hydrogen atoms with a single bond. Thus, “mutants” are similar to, but distinct from, other population members (Figure 21B).

The AutoGrow mutation operator draws upon fragment libraries in order to make small modifications to evolving ligands. By default, AutoGrow comes with large- and small-fragment libraries, although customized, user-generated libraries are straightforward to incorporate. When larger refinements are desired, the mutation operator can draw upon the default library of large molecular fragments. To generate the large-fragment library, 32,091 compounds were downloaded from the ZINC database.⁸⁵ Those ligands with more than 19 atoms were discarded, thereby preserving

only 2,437 small, fragment-like molecules ranging in size from 8 to 19 atoms (average: 16.7 atoms, standard deviation: 2.1, Figure 22a). Collectively, these fragment-like molecules constitute the default large-fragment library. Ten representative large fragments from the library are shown in Figure 23a.

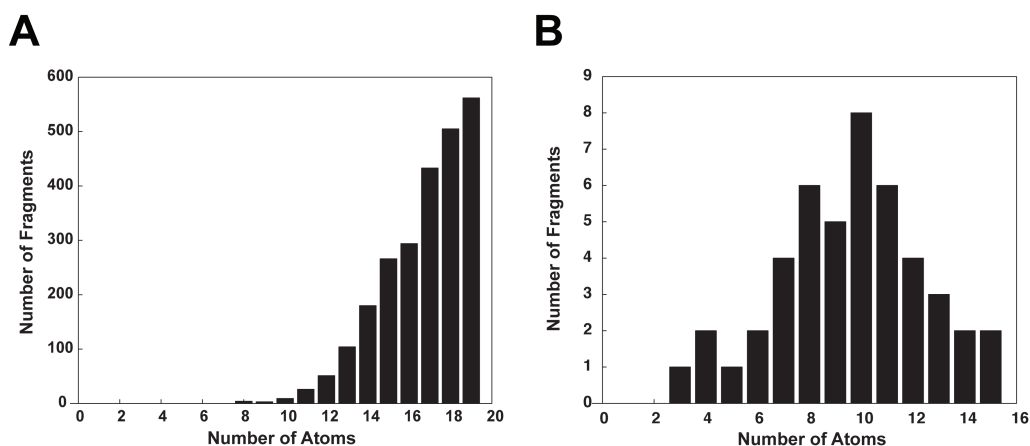


Figure 22: Default large- and small-fragment libraries used by the mutation operator. A) The large-fragment library contains 2,437 fragments ranging in size from 8 to 19 atoms (average: 16.7 atoms, standard deviation: 2.1). B) The small-fragment library contains 46 fragments ranging in size from three to fifteen atoms (average: 9.6 atoms, standard deviation: 2.8).

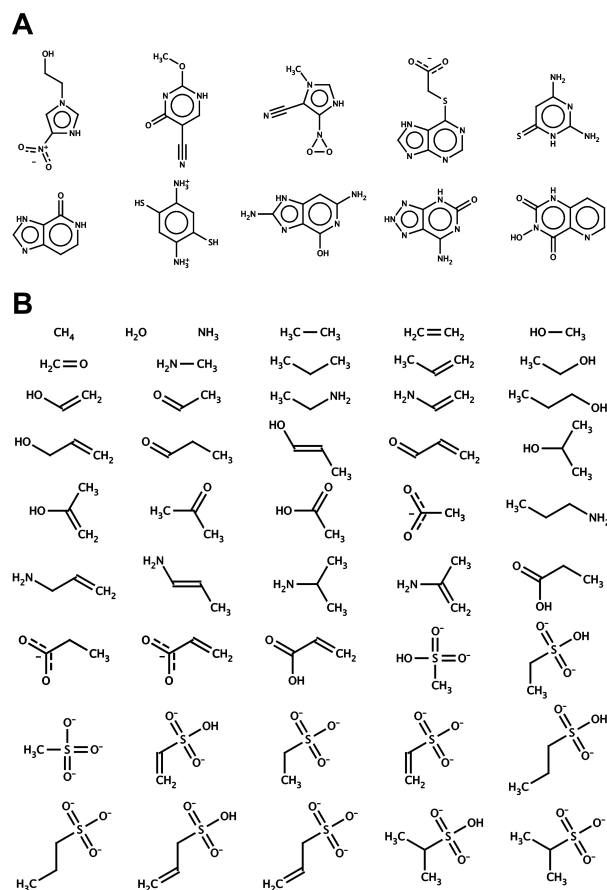


Figure 23: The large- and small-fragment libraries. A) Ten representative fragments were selected from the large-fragment library, which contains 2,437 fragments total. B) All 46 molecular fragments of the small-fragment library.

When small refinements are desired, the mutation operator can draw upon the default library of smaller fragments. To generate the small-fragment library, 46 fragments were created manually using the program *molden*.²²⁹ Most of these small fragments contained one to three carbon atoms, connected by varying combinations of single and double bonds. To these carbon backbones amino, hydroxyl, sulfonate, sulfonic-acid, carboxylate, carboxyl, and ketone moieties were added. Additionally, ammonia and water were included in the library. The size of the small fragments

ranged from three to fifteen atoms (average: 9.6 atoms, standard deviation: 2.8, Figure 22b). All 46 fragments of the small-fragment library are shown in Figure 23b.

To control the AutoGrow algorithm, user-specified variables must be included in a user-input file. A number of the AutoGrow variables in the user-input file correspond to AutoGrid and AutoDock parameters defined in a typical AutoGrid grid-parameter file (*gpf*) or AutoDock docking-parameter file (*dpf*), respectively. Although AutoGrow provides default AutoGrid and AutoDock parameters, these values are receptor dependent and should be modified as necessary. The AutoGrow “*best dock criteria*” variable allows the user to specify whether AutoGrow should judge the fitness of a given ligand based on the AutoDock-predicted binding energy of its lowest-energy or most-populated cluster. In addition to specifying the parameters that AutoGrow will pass to the AutoDock and AutoGrid programs, the user must also specify the parameters of the AutoGrow evolutionary algorithm itself. These parameters include the number of best-fit ligands from each generation to become the founding members of the next generation, as well as the number of “children” and “mutant” ligands derived from those founders via the “crossover” and “mutation” operator, respectively.

The user can place a number of constraints on the growing ligands. User-defined variables specifying the maximum number of generations for which AutoGrow will run as well as the maximum number of atoms allowed per ligand can be used to prevent ligands from growing too large. The user can also require that the

evolving ligands dock at a specific location within the active site and that fragments be added only to selected “core” scaffold hydrogen atoms.

Ligand and Receptor Preparation/Docking Parameters for the Validation Studies

For each of the three validation studies, the receptor proteins were processed prior to AutoGrow execution. Optimized AutoGrid and AutoDock parameters for each of the systems were either taken from the literature or determined here.

For the *TbREL1*-ATP complex, the protein was processed as described by Amaro et. al.³¹ To use AutoGrow to regenerate the ATP substrate, we first identified AutoDock parameters that could successfully redock the co-crystallized ATP molecule back into the 1XDN structure when three crystallographic active-site water molecules essential for proper ATP binding were retained. These AutoDock parameters, tailored specifically for 1XDN, were used in all subsequent AutoGrow-guided AutoDock runs. Decomposing the co-crystallized ATP ligand and adding hydrogen atoms as necessary generated three ATP-derived molecules. Two of these, a triphosphate with a terminal hydroxyl group and an adenine, constituted a fragment library. The third, (3R,4S)-5-(Methyl)tetrahydrofuran-3,4-diol, served as the initial “core” scaffold. Fragments were allowed to grow from four scaffold linker hydrogen atoms: the three equivalent hydrogen atoms of the methyl group and the 5C hydrogen *trans* to the two hydroxyl groups of the tetrahydrofuran ring (Figure 24a).

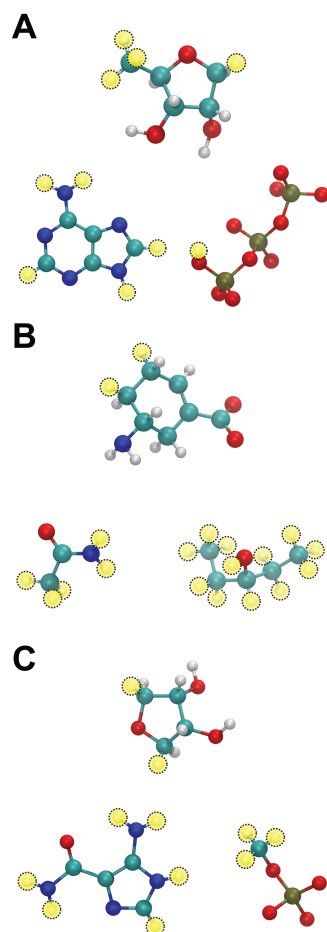


Figure 24: New fragment libraries were created to see if AutoGrow could recreate (A) ATP, (B) oseltamivir, and (C) AMZ. In all cases, the initial “core” scaffold is shown above the two fragments of each library, and yellow circles depict linker hydrogen atoms. A) To regenerate ATP, a library was created containing only two fragments: a triphosphate with a terminal hydroxyl group and an adenine. (3R,4S)-5-(Methyl)tetrahydrofuran-3,4-diol served as the initial scaffold, and fragments were allowed to grow from four scaffold hydrogen atoms. B) To regenerate oseltamivir, a library was created containing only two fragments: pentan-3-ol and acetamide. (5R)-5-aminocyclohex-1-ene-1-carboxylate served as the scaffold, and fragments were allowed to grow from two scaffold linker hydrogen atoms. C) To regenerate AMZ, a library was created containing only two fragments: methoxy(oxo)phosphinate and 5-amino-1H-imidazole-4-carboxamide. (3R,4S)-oxolane-3,4-diol served as the initial scaffold, and fragments were allowed to grow from two scaffold linker hydrogen atoms.

For the neuraminidase-oseltamivir complex, oseltamivir was extracted from the active site, and hydrogen atoms were added using the PRODRG server.²³⁰ AutoDockTools (ADT) version 1.5.2²³¹ was then used to remove non-polar ligand hydrogen atoms, to add protein-receptor polar hydrogen atoms, and to assign Gasteiger charges²³² to all atoms of the system. For this system, we identified AutoDock parameters that could successfully redock the co-crystallized oseltamivir molecule back into the 2HU4 structure, similar to parameters reported in the literature.¹⁰⁹ These optimized AutoDock parameters were used in all subsequent AutoGrow-guided AutoDock runs. The co-crystallized oseltamivir molecule was decomposed into three oseltamivir-derived molecules. Two of these, pentan-3-ol and acetamide, comprised the fragment library. The third, (5R)-5-aminocyclohex-1-ene-1-carboxylate, served as the initial “core” scaffold. Fragments were allowed to grow from two scaffold linker hydrogen atoms (Figure 24b).

For the AICAR transformylase-AMZ complex, we first removed from the crystal structure the AMZ ligand, all but six crystallographic water molecules judged important for AMZ binding, and all potassium ions.²³³ Xanthosine-5'-monophosphate (XMP) and N-[(s)-(4-[(2-amino-4-hydroxyquinazolin-6-yl)(dihydroxy)-lambda~4~-sulfanyl]amino}phenyl)(hydroxy)methyl]-l-glutamic acid (compound 354), two ligands that were co-crystallized with AMZ, were retained as part of the receptor structure. As ADT 1.5.2 does not assign hydrogen atoms to water molecules, the system was processed with ICM (Molsoft), which added hydrogen atoms to the protein and to the water oxygen atoms, appropriately protonated histidines, flipped

certain protein amide groups, and corrected terminal phase angles. Hydrogen atoms were added to the AMZ ligand with Discovery Studio (Accelrys). ADT²³¹ was then used to remove nonpolar hydrogens and assign Gasteiger charges²³² to both the protein and the ligand structures. We then generated new optimized AutoDock parameters that could successfully redock the co-crystallized AMZ molecule back into the 1P4R structure. These tailored AutoDock parameters were used in all subsequent AutoGrow-guided 1P4R AutoDock runs. The co-crystallized AMZ ligand was decomposed into three AMZ-derived ligands; hydrogens were added to these ligands using the PRODRG server.²³⁰ Two of these new molecules, methoxy(oxo)phosphinate and 5-amino-1H-imidazole-4-carboxamide, constituted the fragment library. The third, (3R,4S)-oxolane-3,4-diol, served as the initial “core” scaffold. Fragments were allowed to grow from two scaffold linker hydrogen atoms (Figure 24c).

AutoGrow Parameters

For each of the three test systems, we ran AutoGrow using an initial generation size of fifty compounds. Each subsequent generation also contained fifty compounds. Ten of those compounds were taken from the previous generation, based on both the score of the best AutoDock cluster (either the most-populated or the lowest-energy cluster, depending on the system) and successful active-site binding. An additional twenty “children” and twenty “mutants” were created from these ten primary individuals, subject to the requirement that all compounds contain fewer than seventy atoms. The first generation initially contained only the scaffold and 49 “mutants,” as

no previous generation existed from which “parents” could be drawn for crossover production.

AutoGrid version 4.0 was used to create affinity grids centered on the active site of each protein receptor. Affinity grids were calculated for all of the atom types present in the target ligand (ATP, oseltamivir, and AMZ, respectively). AutoGrow docked all generated ligands into their corresponding protein receptors with AutoDock 4.0.1.²⁶

8.3 Results/Discussion

In the current work, we introduce AutoGrow, a new computer-aided drug design algorithm that uses a growing strategy to build upon an initial “core” scaffold. Molecular fragments are added at random to this scaffold, thereby generating a population of novel ligands. These ligands are subsequently docked into the target protein receptor. An evolutionary algorithm evaluates the docking scores of each population member, and the best binders become founders of the subsequent generation. As generation after generation is created, each based on the most fit individuals of the previous generation, a larger inhibitor with higher predicted binding affinity eventually evolves.

Method Validation

In order to validate the AutoGrow algorithm, we attempted to recreate several known small-molecule protein ligands, including ATP, the natural substrate of *Trypanosoma brucei* RNA editing ligase 1 (*TbREL1*); oseltamivir, a nanomolar-affinity antiviral inhibitor of the influenza neuraminidase enzyme; and aminoimidazole 4-carboxamide ribonucleotide (AMZ), a nanomolar-affinity inhibitor of AICAR transformylase.

AutoGrow limits combinatorial explosion by drawing upon fragment libraries and using an evolutionary algorithm. As a result, AutoGrow often generates compounds with higher predicted binding affinities than the initial scaffold, despite the fact that most modifications to a good scaffold reduce predicted binding affinity. However, the use of even a moderately sized fragment library leads to a computationally intractable number of possible outcomes (ligands); consequently, AutoGrow is nondeterministic. If the goal is to produce a ligand that has a higher predicted binding affinity than an initial scaffold, a solution can be reasonably expected. If the goal is to produce one *specific* ligand out of the millions of possible compound outcomes, success is unlikely.

In order to have any reasonable chance of recreating a specific compound such as ATP, oseltamivir, or AMZ, the number of possible combinations must be significantly reduced. To achieve this reduction, instead of using the AutoGrow default large- or small-fragment libraries, we cut each of these specific compounds into three pieces, producing two fragments and one initial “core” scaffold for each.

Fragments were allowed to grow from appropriate *scaffold* linker hydrogens, though all fragment hydrogen atoms were allowed to serve as linkers (Figure 24). Despite these limitations imposed on the fragment-library size and the number of scaffold linker hydrogen atoms, this setup can still produce hundreds of distinct compounds after only several generations.

Validation Using ATP

ATP, the natural substrate of *TbREL1*, is a good candidate for validation, as the ATP-bound complex has been crystallized at 1.2 Å resolution.⁸⁶ AutoGrow created ATP in the second generation via mutation by adding the adenine fragment to the second best ligand of the first generation; ATP was the highest scoring ligand of the second generation. The ATP adenosine moieties of both the co-crystallized and docked ligand sit deep within the binding site and interact with the protein via a hydrogen bond with E86 and via pi-pi interactions with F209. Additionally, the N1 adenine nitrogen accepts a hydrogen bond from a co-crystallized water molecule. Two conserved residues, E159 and N92, interact with the ATP ribose moiety via hydrogen bonds.

The co-crystallized and AutoGrow-generated ligands do not overlap in the region of the ATP polyphosphate tail. A magnesium ion, present in the crystal structure but absent in the protein receptor used for docking, coordinates many of the polyphosphate oxygen atoms; in the absence of magnesium, the position of the polyphosphate tail is naturally altered (Figure 25). The RMSD between the co-

crystallized ATP and the docked, AutoGrow-generated ATP with all hydrogen atoms removed is 3.88 Å. When the polyphosphate tail is removed (neglecting all atoms including and beyond the alpha phosphate), the RMSD difference falls to 0.39 Å.

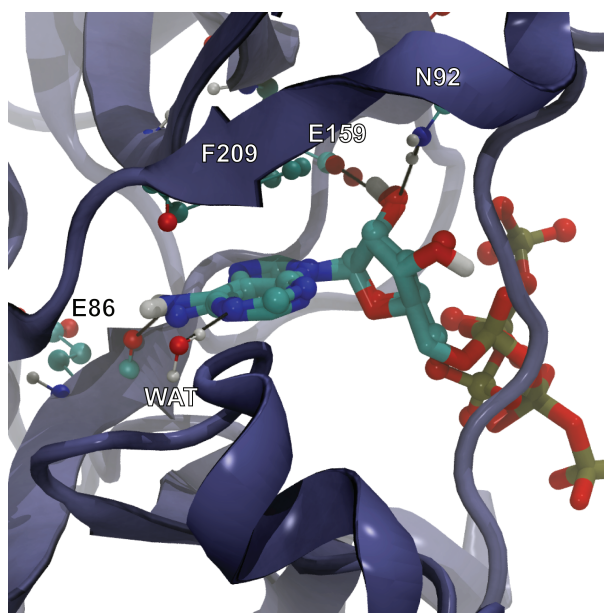


Figure 25: ATP co-crystallized with *TbREL1* (rendered in CPK) superimposed on the AutoGrow-generated ATP docked into the protein receptor (rendered in licorice). Black lines represent hydrogen bonds. The co-crystallized and docked ligands do not overlap in the region of the ATP polyphosphate tail because a magnesium ion, present in the crystal structure but absent in the protein receptor used for docking, coordinates many of the polyphosphate oxygen atoms.

Validation Using Oseltamivir

Oseltamivir is a nanomolar-affinity antiviral compound that inhibits the neuraminidase enzyme; the oseltamivir-neuraminidase (N1) complex has been crystallized at 2.50 Å resolution.¹¹¹ Although AutoGrow failed to recreate oseltamivir after eight generations, inspection of the final set of compounds revealed that AutoGrow had created a very similar compound (Figure 26). This compound, the third best ligand of the second generation, was created as a result of a crossover between the

eighth and ninth best ligands of the first generation. Like oseltamivir, the new compound is predicted to make hydrogen bond interactions with the side chains of R292, R121, E119, D151, and R152 (Figure 26). However, the linking amide nitrogen of oseltamivir, a potential hydrogen bond acceptor, is absent in the new compound. Interestingly, the crystal structure of the neuraminidase-oseltamivir complex demonstrates that this nitrogen does not participate in hydrogen bonding.¹¹¹ In the new compound, the amide nitrogen is further away from the central ring, allowing it to potentially form hydrogen bonds with the backbone carbonyls of W178 and D151. The RMSD between the co-crystallized oseltamivir and the docked, AutoGrow-generated novel compound with all hydrogen atoms removed is 1.95 Å, suggesting that even with the altered acetamide moiety, the binding of the novel compound closely mimics that of oseltamivir (Figure 26).

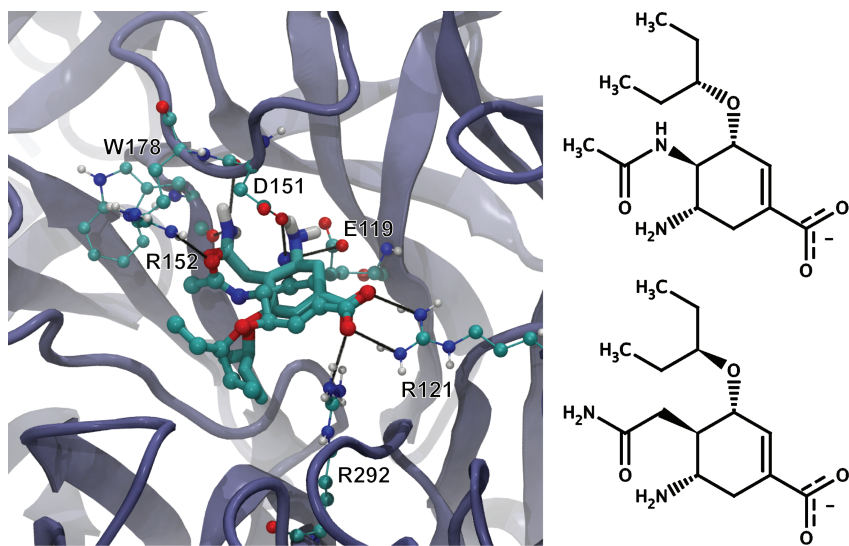


Figure 26: On the left, oseltamivir co-crystallized with neuraminidase (rendered in CPK) superimposed on the AutoGrow-generated novel compound docked into the protein receptor using AutoDock 4.0 (rendered in licorice). Black lines represent hydrogen bonds. On the right, a two dimensional representation of oseltamivir and the AutoGrow-generated novel compound demonstrates their similarity.

Validation Using AMZ

AMZ is a nanomolar-affinity inhibitor of AICAR transformylase; the AMZ-protein complex has been crystallized at 2.55 Å.²²⁶ AutoGrow was able to recreate AMZ in the third generation via mutation by adding the 5-amino-1H-imidazole-4-carboxamide fragment to the sixth best ligand of the second generation; AMZ was the highest scoring ligand of the third generation.

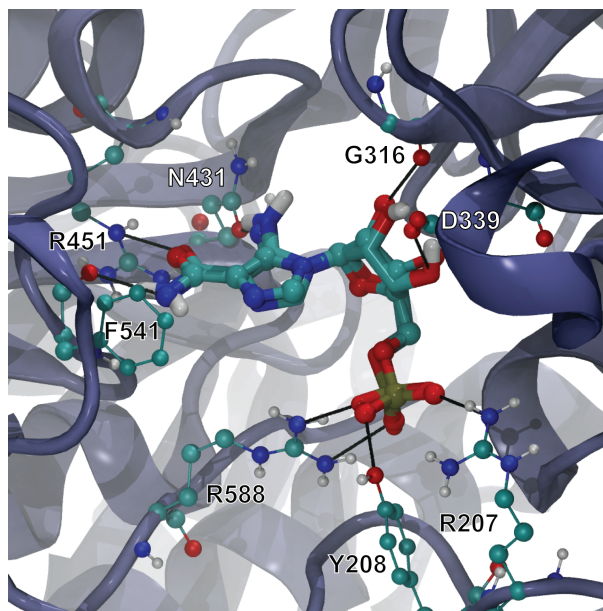


Figure 27: AMZ co-crystallized with AICAR transformylase (rendered in CPK) superimposed on the AutoGrow-generated AMZ docked into the protein receptor (rendered in licorice). Black lines represent hydrogen bonds. For clarity, water molecules present in the active site, which are essential for AMZ binding, are not shown.

The AutoDock-predicted binding pose of the AutoGrow-generated AMZ ligand was nearly identical to the crystallographic pose, with a heavy-atom RMSD of only 0.48 Å (Figure 27). The crystal structure and predicted docked pose of AMZ bound to AICAR transformylase reveal a number of protein-ligand and water-ligand hydrogen bonds. The hydroxyl group of the Y208 side chain and the guanidine groups of the R588 and R207 side chains all make hydrogen bonds with the non-bridging phosphates of the AMZ phosphinate moiety, as do four co-crystallized water molecules buried within the active site. The carboxylate group of the D339 side chain and the G316 backbone carbonyl accept hydrogen bonds from the two hydroxyl groups of the central (3R,4S)-oxolane-3,4-diol core, and two additional co-crystallized water molecules donate hydrogen bonds to the oxygen atoms of those same groups.

Additionally, the carbonyl oxygen of the N431 side chain accepts a hydrogen bond from the amino group attached to the AMZ imidazole ring, and the benzene ring of compound 354 co-crystallized with AMZ stabilizes that imidazole via pi-pi interactions (not shown in Figure 27). Finally, the guanidine group of the R451 side chain donates two hydrogen bonds to the amide carbonyl oxygen of AMZ, and the F541 backbone carbonyl accepts a hydrogen bond from the AMZ amide nitrogen.

Advantages and Limitations of the Current Method

AutoGrow has a number of advantages over other *de novo* drug design algorithms. AutoGrow is novel in the way it combines an evolutionary algorithm with a fragment-based growing method to explore new chemical space and a fitness function based on a fully flexible ligand docking score. Several other programs also combine fragment-based growing with an evolutionary algorithm, including Chemical Genesis,²³⁴ LEA,²³⁵ and LigBuilder,²³⁶ but these algorithms do not incorporate full flexible-ligand docking into their respective fitness functions. AutoGrow performs a separate docking of the entire molecule (moieties and scaffold) for each ligand it creates, allowing the scaffold position to change in response to the addition of any new fragments. Although the *ab initio* redockings increase the computational expense, in many cases they allow for a more realistic prediction of the ligand binding pose.

Two programs, ADAPT²³⁷ and SYNOPSIS,²³⁸ are similar to AutoGrow in that they are fragment-based evolutionary algorithms that do incorporate docking scores into the fitness function. ADAPT uses DOCK,²³⁷ and SYNOPSIS is designed to work

with any number of user-generated fitness functions. For the latter, Vinkers et al. gave one example where an in house docking-program score was used to determine fitness.²³⁸ Despite these similarities, however, AutoGrow differs from ADAPT and SYNOPSIS in the way its evolutionary algorithm is implemented. ADAPT and AutoGrow have similar crossover operators, but very different mutation operators. While ADAPT generates its initial population of ligands by adding fragments to a core scaffold, in subsequent generations it “mutates” ligands by either changing atom types or altering local connectivity. AutoGrow “mutates” ligands by adding novel fragments at each generation, not just the first. In contrast, SYNOPSIS and AutoGrow have somewhat similar mutation operators. However, SYNOPSIS has no crossover operator whereby two ligands can be combined to produce a third with chemical properties of the two “parents.”

AutoGrow has several limitations. As AutoDock is used as the selection operator, the results are dependent on the accuracy of the AutoDock scoring function, which has a standard error of 2.177 kcal mol⁻¹.²⁶ The AutoDock scoring function used to predict ligand free energies of binding is much faster and less accurate than comprehensive physics-based approaches like thermodynamic integration,⁹¹ single-step perturbation,⁹² and free energy of perturbation.⁹³ Nevertheless, a recent validation study⁹⁴ showed that AutoDock performs well compared to other common docking programs like DOCK,⁹⁵ Flex,⁹⁶ and GOLD,⁹⁷ and AutoDock is the most cited docking software in the literature.²²⁸ Additionally, unlike many other docking programs, AutoDock is freely distributed.

AutoDock employs a Lamarckian genetic algorithm that is computationally intensive and time consuming; one study found that AutoDock was slower than DOCK, Flex, GOLD, and ICM.⁹⁴ AutoGrow generates novel compounds quickly, but its dependence on AutoDock impacts the efficiency of the algorithm. The extent of impact depends on the exact AutoDock parameters used and the computer hardware employed. As an example, an analysis of the AutoGrow data from one of our own ongoing projects shows that AutoDock docked each of our AutoGrow-generated ligands in 4.1 hrs $\pm$ 1.9. Running on 51 processors, each generation of 50 ligands finished in 6.5 hrs $\pm$ 2.6.

Inaccuracies in the predicted binding energies may arise because AutoDock only approximates ligand flexibility *via* torsional rotation and molecule translation and rotation. All bond lengths are fixed so ligand flexibility is not truly complete; however, the effects due to changes in bond length are expected to be minor. Additionally, AutoDock uses an implicit desolvation term that is not as accurate as more computationally expensive explicitly solvated molecular dynamics techniques.

Although AutoGrow is not presently coded to interface with alternate docking programs, therefore precluding an explicit comparison, we anticipate that incorporating a faster docking and scoring module will improve the computational performance. Other methods to improve performance, such as incorporating a parallelized version of AutoDock,²³⁹ can also be envisioned. As AutoDock is free, performs well compared to many other docking programs,⁹⁴ and has a well-tested scoring function,²²⁸ AutoGrow has been coded to interface with AutoDock by default;

with only slight modification to the code, however, AutoGrow could interface with any number of docking programs. We note that AutoGrow is open source, and we encourage source-code modifications of this type.

Other AutoGrow limitations are typical of any computer-aided drug design method. With the number of chemically feasible, drug-like molecules being on the order of 10^{60} to 10^{100} ,⁴⁴ the size of chemical space is far too large to be completely scanned by any computer algorithm. While AutoGrow can in many cases generate compounds predicted to bind with high affinity, one cannot expect it to find the single best binder. As evolutionary algorithms are often nondeterministic, we recommend running AutoGrow on the same initial scaffold several times in order to obtain multiple potential inhibitor candidates for subsequent evaluation and optimization.

The quality of the AutoGrow results is also affected by the fragment database used. Medicinal chemists should carefully filter AutoGrow results, eliminating or modifying metabolically reactive groups and other undesirable moieties. AutoGrow is unlikely to produce good *de novo* drug candidates if the default fragment databases are used. The program does not take into account Lipinski's Rule of Five⁷⁵ or ADME-Tox properties. Consequently, one should not expect AutoGrow-generated compounds to be drug-like. However, the compounds that are generated can help the medicinal chemist identify novel and potentially important receptor-ligand interactions. AutoGrow-suggested compounds may also elucidate new drug scaffolds, thereby providing useful information during the initial stages of drug development. We

recommend using AutoGrow to supplement the medicinal chemist's creativity, rather than to replace it.

AutoGrow can be downloaded for free at <http://autogrow.ucsd.edu>

8.4 Acknowledgements

REA is funded by NIH F32-GM077729 and NSF MRAC CHE060073N. Funding by NIH GM31749, NSF MCB-0506593 and MCA93S013 (to JAM) also supported this work. Additional support from the Howard Hughes Medical Institute, the National Center for Supercomputing Applications, the San Diego Supercomputing Center, the W.M. Keck Foundation, Accelrys, Inc., the National Biomedical Computational Resource and the Center for Theoretical Biological Physics is gratefully acknowledged. We would also like to thank Meghan Cain for helpful discussions and code debugging.

Chapter 8, in full, is a near reprint of the material as it appears in Chemical Biology and Drug Design 2009, volume 73, pages 168-178. The dissertation author was the primary contributor to this work. Rommie E. Amaro also made significant contributions, and J. Andrew McCammon served as head adviser and principal investigator. Jacob Durrant holds the copyright for this manuscript and has authorized its publication in this dissertation.

CHAPTER 9: Including Receptor Flexibility and Induced Fit Effects into the Design of MMP-2 Inhibitors

Jacob D. Durrant, César Augusto F. de Oliveira, and J. Andrew McCammon

9.1 Introduction

Matrix metalloproteinases (MMPs) comprise a class of flexible proteins²⁴⁰⁻²⁴⁴ for which computational drug design has been very challenging.²⁴⁵ There are over 20 MMP enzymes²⁴⁶ capable of degrading nearly all the components of the extracellular matrix, as well as many other extracellular proteins.^{22, 247-250} MMPs are required for normal tissue remodeling and are involved in many physiological processes, including embryonic growth, signal regulation, inflammation, angiogenesis,^{246, 251} uterine involution, bone resorption, wound healing,^{22, 248-250} CNS regeneration,²⁵² and ovulation.²⁵³

Overexpression of MMPs has been associated with a wide range of pathophysiological processes as well, including vascular disease,²⁵¹ asthma,²⁵⁴⁻²⁵⁶ arthritis, multiple sclerosis, Alzheimer's disease,²⁵⁷ and cancer,²⁵⁸⁻²⁶⁰ which requires degradation of the extracellular matrix for tumor progression and metastasis. The development of MMP inhibitors for the prevention of cancer metastasis has been a particularly active field of study.²⁶¹ MMP-2 (gelatinase A) and other gelatinases have received special attention because they are most consistently detected in malignant tissues and are associated with tumor aggressiveness, metastatic potential, and poor prognosis.^{261, 262} Nearly all MMP inhibitors have failed in clinical trials, in part due to lack of specificity.²⁶³

Any rational design of MMP-2 inhibitors must take into account the general shape and characteristics of the MMP-2 active site.²⁶³ MMP inhibitors are typically comprised of a zinc-binding group (ZBG) with attached moieties that interact with surrounding binding pockets. A number of zinc-chelating groups have been proposed in the literature, including hydroxamates,²⁵³ thiols,²⁶⁴ pyrone derivatives,²⁶⁵⁻²⁶⁷ carbamoylphosphonates,^{268, 269} hydroxyureas,^{270, 271} hydrazines,²⁷² β -lactams and squaric acids,^{273, 274} and nitrogenous ligands (dipicolylamine).^{275, 276} Among these chelators, hydroxamic acid is by far the most frequently used.^{253, 264}

Cohen et al. recently used a bioinorganic approach to discover new classes of zinc-chelating compounds that show improved potency and novel selectivity relative to similar hydroxamate-based ZBGs.^{265, 275, 277} These ZBGs were crystallized bound to a hydrotris(3,5-phenylmethylpyrazolyl)borate bioinorganic model complex whose zinc-coordinating nitrogen atoms are geometrically positioned in a way that is analogous to the three zinc-coordinating histidine nitrogen atoms of the MMPs,²⁶⁶ allowing the ZBG to be modeled within any MMP-2 active site *in silico* by aligning analogous model and protein atoms.

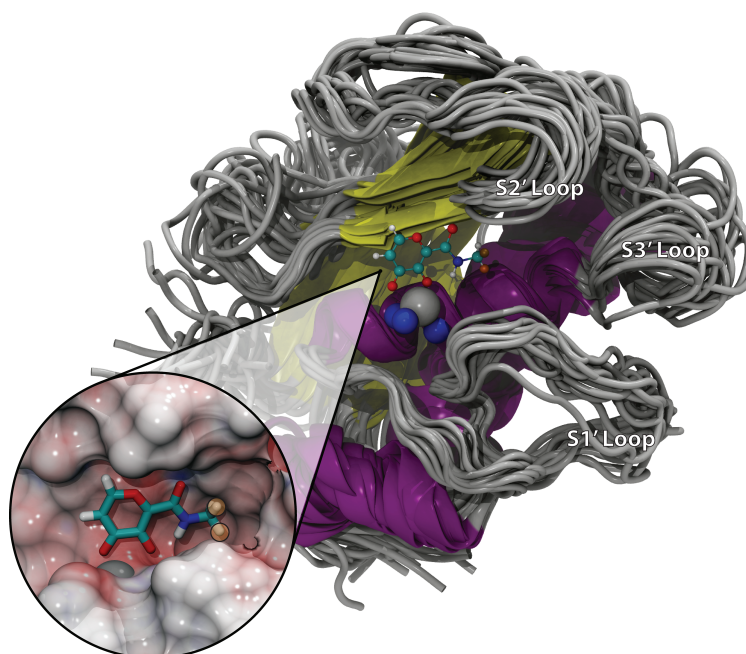


Figure 28: Superposition of equally spaced snapshots extracted from the MD trajectory of the *apo* protein. Fitting was performed by aligning the catalytic zinc cation and histidine residues of each structure. The zinc cation (grey) and coordinating nitrogen atoms of histidine sidechains (blue) are shown in van der Waals representation. The ZBG, maltol (3-hydroxy-2-methyl-4-pyrone), and an amide link are shown in balls and sticks representation. The inset figure shows the ZBG (licorice representation) and the catalytic zinc cation (van der Waals representation) in the MMP-2 active site (surface representation colored according to the protein electrostatic potential). The yellow circles highlight the hydrogen atoms used as LUDI linking points through which interacting fragments were added.

One identified ZBG, 3-hydroxy-2-methyl-4-pyrone (maltol), is amenable to the addition of moieties that simultaneously interact with multiple MMP binding pockets. This is especially true when maltol is modified to 3-hydroxy-*N*-methyl-4-oxo-4*H*-pyran-2-carboxamide. Molecular modeling suggests that two of the hydrogen atoms on the terminal methyl group of this modified maltol point towards the S1' and S2' pockets, respectively (Figure 28). S1' is a cavernous, flexible, hydrophobic pocket that is important for substrate selectivity.²⁶³ The S2' pocket has the greatest geometric

variability of all sub-pockets, as judged by force-field interaction energies used to compare the structures of 24 MMPs (nine X-ray structures and 15 homology models). As the shape of the S2' pocket is not well conserved across the many different MMPs, targeting this pocket may lead to better therapeutic specificity.²⁷⁸

Table 11: MMP-2 inhibitor candidates obtained by applying the procedure illustrated in Figure 29.

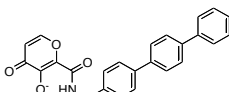
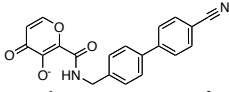
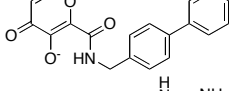
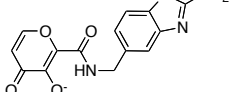
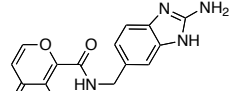
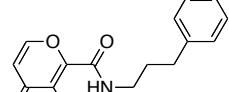
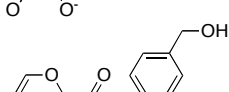
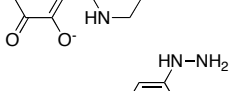
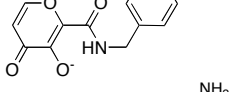
ID	Molecular Structure	LUDI Frequency	LUDI Max Score	Score ^{apo} (kcal/mol ⁻¹)	Score ^{holo} (kcal/mol ⁻¹)	Pocket
AM-6		0.24	716	-12.36	-11.49 ± 0.65	S1'
AM-5		0.86	635	-10.27	-9.86 ± 0.56	S1'
AM-2		0.77	611	-10.22	9.18 ± 0.58	S1'
UD2a		0.52	712	-8.79	-8.20 ± 0.70	S1'
UD2b		0.37	722	-4.40	-7.89 ± 0.70	S1'
T12		0.73	481	-7.11	-7.52 ± 0.63	S1'
UB9		0.63	594	-6.67	-7.05 ± 0.48	S1'
UG1		0.43	749	-7.53	-6.45 ± 0.52	S1'
UF7		0.35	775	-6.86	-6.44 ± 0.82	S1'

Table 11 continued:

ID	Molecular Structure	LUDI Frequency	LUDI Max Score	Score ^{apo} (kcal/mol ⁻¹)	Score ^{holo} (kcal/mol ⁻¹)	Pocket
T12		0.65	435	-6.22	-5.99 ± 0.48	S2'/S3'
AM-6 ^a		x		-9.05	-5.93 ± 0.60	S1'
T02		0.62	398	-4.99	-5.34 ± 0.58	S1'
U97		0.35	576	-6.56	-5.30 ± 0.76	S2'/S3'
UA5		0.39	654	-5.01	-5.04 ± 0.55	S2'/S3'
T03		0.68	416	-5.37	-4.74 ± 0.61	S1'
UE8		0.64	602	-7.53	-4.64 ± 0.64	S1'
T18		0.66	450	-5.10	-4.55 ± 0.76	S1'
U97 + UD2a ^b		-	-	-7.31	-9.28 ± 0.87	Both
U97 + UD2a ^c		-	-	-9.41	-9.20 ± 0.64	Both

^a Based on a more representative LUDI score (450), not the best LUDI score. Associated productive *holo* simulation of only 1.35 ns, not 3.5 ns.

^b Using the optimal U97 protein configuration.

^c Using the optimal UD2a protein configuration.

Due to the high flexibility of the MMP binding pockets,²⁴⁰⁻²⁴⁴ predicting the binding affinity of MMP inhibitors *in silico* is particularly difficult. Although structural information extracted from the several available X-ray structures can be used, in principle, to design and improve the selectivity of a given inhibitor towards a specific MMP, the high flexibility of the MMP binding pockets drastically limits the accurate modeling and prediction of inhibitor-protein complexes.^{279, 280}

In order to assist future drug design, our group recently studied the flexibility of the *apo* MMP-2 binding pockets *via* a molecular dynamics (MD) simulation.²⁸¹ The MD simulation revealed that two states, characterized as open and closed, are preferentially sampled. By analyzing the population of those states along the trajectory, we observed that while the closed state is heavily sampled, the open state, in which the S1' pocket adopts a tunnel-like shape, appears only rarely. The S1' loop, which has the highest sequence variability among different MMPs, is among the most mobile segments of MMP tertiary structure. Loops S2' and S3', which encompass solvent-exposed pockets S2' and S3', are also highly flexible (Figure 28).

Previously, our group presented a novel virtual-screening approach called the “relaxed complex” scheme that accounts for receptor flexibility.²⁹ In this scheme, protein flexibility is taken into account by docking candidate inhibitors into an ensemble of protein configurations extracted from MD simulations of either the *apo* or *holo* receptor.^{28, 282} This approach has been applied to several proteins including HIV-1 integrase,^{30, 283} *Trypanosoma brucei* RNA editing ligase 1,³¹ and influenza neuraminidase.¹⁰⁹

Given the dynamics of the MMP-2 binding pockets observed previously, a relaxed-complex technique was employed in the current study. Protein configurations were extracted from the *apo* MMP-2 simulation and the ZBG was positioned into each frame using *in silico* modeling based on the Cohen bioinorganic model.^{265, 275, 277} LUDI, a fragment-based *de novo* drug-design algorithm,¹¹⁹ was then used to dock fragments into the MMP-2 active site such that the fragments could be easily merged with the ZBG while maintaining interactions with the S1', S2', and S3' pockets across multiple protein configurations extracted from the *apo* MD simulation.

Next, we extended the relaxed complex scheme by performing secondary MD simulations of the *holo* receptor (receptor in complex with each of the thirteen best LUDI-predicted ligands) in order to account for induced-fit effects. These thirteen compounds were then re-ranked according to their ensemble-average predicted binding energy. Re-ranking by the ensemble-average score, which accounts in part for the open-closed dynamics of the active site, significantly reordered the compounds, suggesting that induced-fit effects can have a substantial impact on the computational identification of potential enzyme inhibitors.

9.2 Computational Details

Initial molecular dynamics simulations were performed on a 2.80 Å resolution MMP-2 crystal structure (PDBID: 1QIB²⁸⁴). Crystal waters and inhibitors were removed from the structure. Basic residues, i.e. Arg and Lys, were protonated, and acidic residues, i.e. Asp and Glu, were deprotonated. Except for one histidine residue

located in the structural domain (His 179), which was protonated in the epsilon position, all histidine residues were protonated in the delta position. To maintain the correct orientation of all zinc-chelating histidine residues and to prevent the escape of zinc atoms to the solvent, harmonic restraints were applied on the distances and bending angles defined by the zinc and the nitrogen atoms of the appropriate histidine residues. The partial charges of the histidine imidazole rings and zinc atoms were calculated using the RESP program. The molecular electrostatic potential was calculated at the HF/6-31G* level.

Following the initial system preparation, we applied 500 steps of steepest descent followed by 1000 steps of conjugate gradient minimization. The relaxed structure was then solvated in a cubic box of pre-equilibrated TIP3P water molecules,⁵⁵ which extended 10 Å further than the protein atoms themselves. To neutralize the system, sodium counterions (Na^+) were added to balance the charge of the MMP-2 protein. To relax the waters, the system was then again minimized for 500 of steepest descent followed by 2000 steps of conjugate gradient minimization using constant volume periodic boundaries, with the protein and counterion atoms fixed. To bring the system to the correct density, we carried out an additional 50 ps MD simulation in which only water molecules were allowed to move and *NPT ensemble* ($T = 298\text{K}$, $P=1\text{ bar}$) was applied. To relax the protein in the solvent environment, all coordinates were then optimized by employing 500 steps of steepest descent followed by 1000 steps of conjugate gradient minimization. We then performed a 500 ps MD simulation to heat the system from 0K to 300K using the *NVT ensemble* ($T = 298\text{K}$).

All subsequent simulations were carried out using the *NVT ensemble* ($T = 298\text{K}$) and, except for the internal restraints mentioned before and SHAKE constraints on bonds to hydrogen atoms,²⁸⁵ the coordinates of all atoms were allowed to move freely. A final MD simulation of 30 ns was performed. The first 15 ns were used to equilibrate the system, and the last 15 ns were used as the productive run.

Three thousand and one frames were next extracted at regularly spaced intervals from the 15 ns productive portion of the MD simulation. For each frame, the complex formed between the ZBG and the receptor was built by aligning each frame with the crystal-structure coordinates of maltol (3-hydroxy-2-methyl-4-pyrone) bound to [(TpPh,Me)ZnOH] (TpPh,Me = hydrotris(3,5-phenylmethylpyrazolyl)borate), a complex that models the zinc(II) ion coordination within the MMP active site.^{286, 287} Maltol was positioned within the protein active site by minimizing the RMSD distance between the appropriate zinc and the coordinating nitrogen atoms of the protein active site and those of the bioinorganic model (Figures 2A and 2B).

LUDI was then used to dock potential interacting fragments into the many receptor configurations extracted from the *apo* MD simulation. The terminal methyl hydrogen atoms of the ZBG were selected as the link sites (Figure 28). Initial docking used the default Discovery-Studio LUDI link fragment library (Accelrys) with the following parameters: maximum alignment angle 20° ; maximum alignment RMSD $0.6 \text{ \AA}$; search radius $11 \text{ \AA}$; rotate bonds two at a time; preselect 4.0; minimum separation 3.0; lipophilic density 40; polar density 40; minimum surface 0; link weight 1.0; lipophilic weight 1.0; H-bond weight 1.0; aliphatic aromatic off; reject bifurcated

off; no unpaired polar off; electrostatic check off; minimum score 0; maximum fits 8000; maximum hits all; maximum unfilled cavity 0; energy estimate 1 scoring function¹¹⁹; and best fit.²⁷⁷

Following the LUDI-guided identification of interacting fragments, the fragments were combined with the ZBG and additional MD simulations of 5 ns were carried out on the resulting protein-ligand complexes comprised of the best LUDI-generated ligand candidates and the MMP-2 receptor. An MD protocol similar to that described above was used. Ligand force-field parameters were determined using the antechamber program inside the AMBER package. Ligand partial charges were calculated using the AM1-BCC charge scheme.²⁸⁸ In order to guarantee that the zinc-chelating group of each candidate ligand remain bound to the zinc cation of the active site, harmonic restraints were applied to the distances between the two ligand chelating oxygen atoms and the catalytic zinc cation.

For all MD simulations, the equations of motion were integrated every 2 fs using the Verlet Leapfrog algorithm.²⁸⁹ During MD, temperature and pressure were controlled *via* a weak coupling to external bath with a coupling constant of 0.5 and 1 ps, respectively.²⁹⁰ The center-of-mass motion was removed at regular intervals of 5 ps. The PME summation method was used to treat the long-range electrostatic interactions in the minimization and simulation steps of the solvated systems.²⁹¹ The short-range nonbonded interactions were truncated using an 8 Å cutoff and the nonbonded pair list was updated every 10 steps. All calculations were performed in the AMBER package¹⁶⁹ with the ff99 force field.²²⁵

In order to estimate the ensemble-average binding energy of each protein-ligand complex, seventy frames were extracted at regular intervals from the last 3.5 ns of each 5 ns *holo* simulation. The AutoDock 4.0 scoring function was used to estimate the ligand-protein binding energy of each frame. AutoDockTools (ADT) version 1.5.2²³¹ was used to remove non-polar receptor and ligand hydrogen atoms and to assign Gasteiger charges²³² to all the atoms of each frame. The MMP-2 zinc cations were manually assigned charges of +2.0 *e*. AutoGrid 4.0 was used to generate affinity grids centered on the active site. Each grid enclosed a volume of 48.00 Å x 48.00 Å x 48.00 Å with 0.375 Å spacing. For each protein-ligand system, all corresponding AutoDock-predicted binding energies were averaged to give the ensemble-average predicted binding energy (Score^{*holo*}).

For all controls, a procedure similar to the above was used, except that the ZBG 3-hydroxy-4-oxo-4*H*-pyran-2-carboxamide was used instead of 3-hydroxy-*N*-methyl-4-oxo-4*H*-pyran-2-carboxamide, and an extra methyl group was placed on the docked fragments to compensate.

9.3 Results and Discussions

Accounting for protein flexibility

The methodology presented here identifies potential inhibitors in a way that accounts for protein flexibility. The approach presented in this work, illustrated in Figure 29, aims to identify ligand candidates that not only interact favorably with

conformations typical of the *apo* form of the protein, but also with potential novel configurations induced by ligand binding.

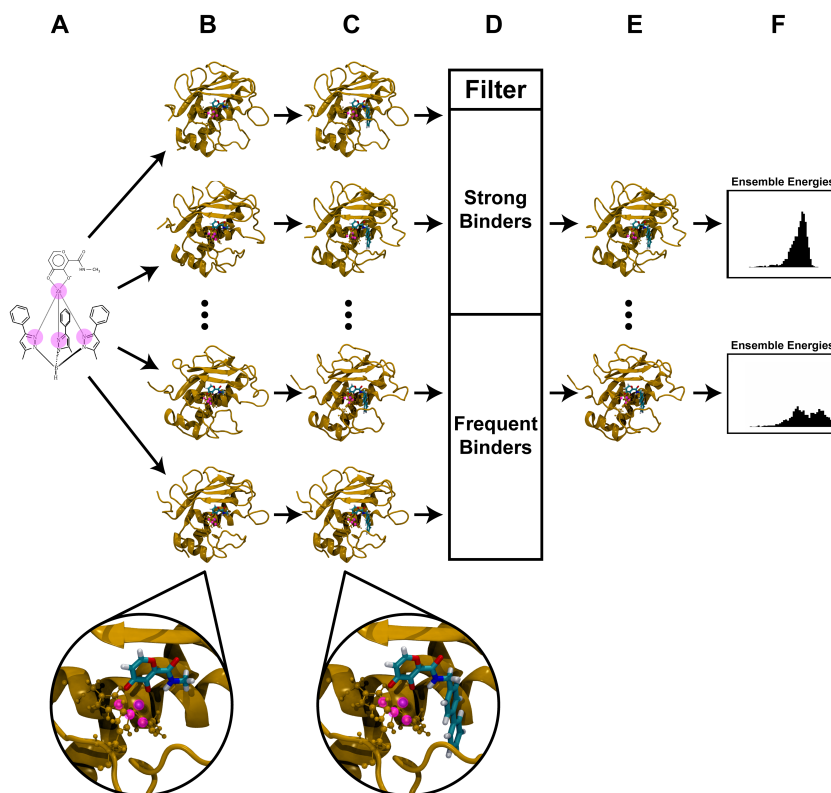


Figure 29: A and B) A molecular dynamics simulation of the *apo* MMP-2 protein was analyzed in order to account for protein flexibility. Three thousand and one regularly spaced frames were extracted from the simulation, and the ZBG was placed into each protein configuration in harmony with the Cohen bioinorganic model. C) LUDI docking calculations were then performed on the 3001 snapshots of the *apo* MD trajectory to add interacting moieties to the ZBG *via* linking hydrogen atoms. D) The “best” interacting fragments were considered to be those that were either frequently selected as binders across multiple protein configurations regardless of LUDI score or had the highest ensemble-wide maximum LUDI score regardless of frequency. E) A 5 ns MD simulation was then performed on each protein-ligand system selected previously. F) To predict the binding affinity of each ligand, frames were extracted at regular intervals from the last 3.5 ns of each of these secondary MD simulations. For each system, the AutoDock 4.0 scoring function was used to estimate the ligand binding energy for all frames; these predicted binding energies were then averaged, yielding an ensemble-average predicted binding energy for each ligand. A small part of this graphic was obtained from reference 287.

Based on the ideas behind the relaxed-complex scheme,^{28, 29, 282} we first performed a molecular dynamics simulation of the *apo* protein in order to account for protein flexibility. Three thousand and one regularly spaced frames sampling both open and closed MMP-2 conformations were extracted from the last 15 ns of the 30 ns simulation, and the zinc-binding group (ZBG, 3-hydroxy-*N*-methyl-4-oxo-4*H*-pyran-2-carboxamide) was manually placed into each protein configuration *via* structural alignment of the protein structure with a crystal-structure bioinorganic model of the MMP-2 active site^{265, 275, 277} (Figures 2A and 2B).

Next, we selected two strategically positioned ZBG hydrogen atoms as fragment linking points (Figure 28, inset picture). The first hydrogen atom is generally oriented toward the S1' pocket and the second is located in the vicinity of the S2' and S3' pockets. By targeting different binding pockets, we aim to identify fragments that simultaneously interact favorably and specifically with the protein. This approach can be easily extended to enable future combinatorial drug design that achieves high affinity by targeting two or more pockets concurrently.

LUDI docking calculations were then performed on the 3001 snapshots of the *apo* MD trajectory using both linking hydrogen atoms (Figure 29C). Because the best candidates obtained from LUDI docking calculations may change according to the conformation of the receptor, two strategies were applied to select the “best” interacting fragments. The best fragments were considered to be those that I) were frequently selected as binders across multiple protein configurations regardless of

LUDI score or II) had the highest ensemble-wide maximum LUDI score regardless of frequency (Figure 29D). Candidates were considered “frequent binders” if they were selected and docked into at least 60% of the protein configurations, regardless of score. Candidates were considered “strong binders” if their ensemble-maximum LUDI scores fell within 100 units of the highest score of any receptor/fragment combination. Finally, we discarded from our list any fragments that failed to bind in the targeted pockets or that contained long linear aliphatic chains, as compounds with long linear aliphatic chains contain many rotatable bonds and, consequently, large entropic penalties upon binding.

Including Induced-Fit Effects: Rescoring Candidate Ligands

In order to account for the induced-fit changes in protein flexibility that occur upon ligand binding, we performed additional MD simulations of each of the LUDI-generated ligands selected previously in complex with the MMP-2 protein. To model each ligand, we iterated through both the frequent and strong binding fragments and in each case identified the fragment and receptor conformations associated with the ensemble maximum LUDI score. Each candidate fragment, posed in its optimal configuration, was subsequently positioned within the corresponding optimal receptor conformation from the *apo* simulation and merged with the corresponding ZBG.

A 5 ns MD simulation of each protein-ligand system was then performed to simulate induced-fit effects (Figure 29E). For each system, 100 regularly spaced frames were extracted from the MD trajectory. The AutoDock 4.0 scoring function

was used to estimate the ligand energy of binding associated with each frame; visualization of the data revealed that the score converged sufficiently after the first 1.5 ns of each simulation. To predict the binding affinity of each ligand, the AutoDock scores associated with the seventy frames of the last 3.5 ns of each of MD simulations were averaged, yielding an ensemble-average predicted binding energy for each ligand (Figure 29F). Candidate ligands were then re-ranked by this ensemble average. Comparing the ensemble-average AutoDock-predicted binding affinity of each system (Score^{holo}) with the predicted binding affinity of the ligand within the *apo* active site following 5000 steps of steepest-descent minimization (Score^{apo}) reveals that accounting for induced-fit conformational changes significantly alters the predicted binding energies of some compounds (Figure 31).

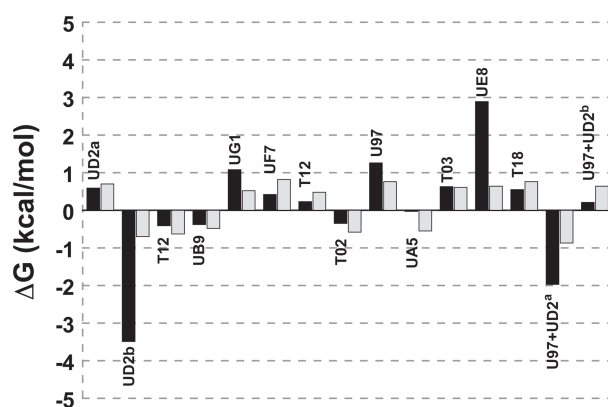


Figure 31: Black) Induced-fit energy contribution. The induced-fit energy contribution is given by $\text{Score}^{apo} - \text{Score}^{holo}$. Grey) Standard deviation of the ensemble-average score. Contributions from induced-fit effects are considered significant if $(\text{Score}^{apo} - \text{Score}^{holo}) > \text{standard deviation}$.

Table 11 shows the potential ligands obtained by applying the procedure depicted in Figure 29. We identified nine fragments predicted to bind in the S1' pocket and three predicted to bind in the S2' or S3' pockets, hereafter referred to jointly as S2'/S3'. As expected, fragments targeting the S1' pocket were generally elongated and hydrophobic. Due to the less hydrophobic and more solvent-exposed characteristics of the S2'/S3' sites, two of the three candidates targeting these pockets were polar and charged. The third, T12 (uncharged), is also a S1' binder; in fact, ensemble-average predicted binding energies suggest T12 prefers the S1' pocket to the S2'/S3' pockets. The predicted binding energies calculated from the *apo* protein (Score^{apo}) were only loosely correlated with the ensemble-average predicted binding energies (Score^{holo}) (R^2 less than 0.3), indicating that altered receptor flexibility due to induced-fit effects has a significant impact on the predicted energy of binding.

For ligands such as those associated with fragments UD2b, UE8, and U97, significant changes in the predicted binding energy occur when the subtle changes in protein configuration that occur upon binding are considered. For instance, the ligand associated with fragment UE8, which was selected as a reliable and strong binder due to its high LUDI frequency and its maximum LUDI score, was among the compounds with the highest binding affinity when only conformations of the *apo* protein were considered (Table 11, see Score^{apo}). However, when induced-fit effects were included in our scoring procedure, the ligand associated with fragment UE8 became the second worst candidate.

Interesting results were also obtained for the ligands associated with the fragments UD2a and UD2b. The high LUDI scores of these tautomers qualified both as strong binders. Nevertheless, after re-scoring the compounds using the AutoDock 4.0 scoring function, the ligand associated with fragment UD2b became the least potent candidate on the list ($\text{Score}^{apo} = -4.40$ kcal/mol), though the ligand corresponding to fragment UD2a still scored well. As expected, after equilibrating each ligand-protein complex, the ensemble-average scores (Score^{holo}) of the tautomers (UD2a and UD2b) converged to nearly the same value.

Table 12: Analysis of active-site structural changes that occur upon ligand binding. The frame from the initial MD simulation that gave the highest LUDI score (“Frame”), the same frame following minimization (“Min”), and the corresponding 3.5 ns productive portion of the associated *holo* trajectory (“Trajectory Average”) were all aligned by minimizing the RMSD between the non-hydrogen atoms of 26 residues lining the active site of the MMP-2 crystal structure (PDB: 1QIB)

ID	Frame (RMSD)	Min (RMSD)	Trajectory Average (RMSD)	$\text{Score}^{apo} - \text{Score}^{holo}$ (kcal/mol)
UD2a	3.57	3.54	3.60 ± 0.08	0.59
UD2b	3.62	3.59	3.62 ± 0.05	-3.49
T12	3.61	3.59	3.37 ± 0.05	-0.41
UB9	3.52	3.51	3.52 ± 0.09	-0.38
UG1	3.50	3.49	3.94 ± 1.15	1.08

To measure the subtle changes in the active-site configuration that yield such drastic changes in the predicted binding energy (i.e. the “induced-fit effects”), we further analyzed the top four predicted inhibitors, those associated with fragments UD2, T12, UB9, and UG1. For each of these inhibitors, the frame from the initial *apo* MD simulation that gave the highest LUDI score, the same frame following minimization with ligand present, and the corresponding 3.5 ns productive portion of

each associated *holo* trajectory were all aligned by minimizing the RMSD between the active-site non-hydrogen protein atoms of 26 residues lining the active site of the MMP-2 crystal structure (PDB: 1QIB, Table 12). The results suggest that even small changes in the active-site configuration, as judged by the small changes in the RMSD to the crystal-structure active site, can lead to substantial changes in predicted binding affinity, as judged by $\text{Score}^{apo} - \text{Score}^{holo}$. Clearly, even limited protein flexibility can have a major impact on the predicted energy of binding.

An analysis of the *holo* trajectories of the top four compounds also reveals a number of common ligand binding modes. The H201 side chain plays an important role in the binding of all these compounds. All four form critical pi-pi interactions with H201; additionally, UD2, UB9, and UG1 also form hydrogen bonds with the H201 side chain. Protein backbone carbonyl oxygen atoms are also critical for binding, as they accept hydrogen bonds from three of the top four ligands; UD2, UB9, and UG1 all donate hydrogen bonds to the backbone carbonyl oxygen atoms of residues I222 and A220. Additionally, UD2 and UG1 also form hydrogen bonds with the backbone carbonyl oxygen atoms of residues L218 and L197. Finally, the backbone carbonyl oxygen atom of P221 and the backbone amine of I222 form hydrogen bonds with UD2 and UG1, respectively.

Controls

In order to validate our methodology, we included three controls, AM-2, AM-5, and AM-6, which are known to inhibit MMP-2 with IC_{50} values of 9.3 μ M, 0.61 μ M, and > 50 μ M, respectively.²⁷⁷ A procedure similar to that used above was repeated with these three controls. Our two positive controls, AM-2 and AM-5, had predicted binding energies of 9.18 ± 0.58 and -9.86 ± 0.56 kcal/mol, respectively. Given that AutoDock has a residual standard error of $2.177 \text{ kcal mol}^{-1}$,²⁶ these positive controls have predicted binding energies comparable to those of the best novel compounds identified (Table 11).

AM-6, included as a negative control, was incorrectly identified as a strong binder (Table 11). In order to understand why the current methodology led to this false positive result, we examined the LUDI dockings of the AM-6 fragment into the protein configurations of the *apo* trajectory. Those dockings with the highest LUDI scores placed the AM-6 fragment deep within the hydrophobic S1' pocket; one of these dockings was subsequently used to construct the initial frame of the *holo* AM-6 MD simulation. However, there were very few of these high-scoring dockings, suggesting that MMP-2 can only rarely accommodate the large AM-6 fragment. This observation is supported in part by the fact that LUDI was only able to dock AM-6 into the MMP-2 active sites of 24% of the frames extracted from the *apo* MD trajectory (Table 11). Even when docking was successful, docking of the AM-6 fragment into a well defined S1' pocket was rare.

After generating a histogram of all AM-6 LUDI scores, we noted that the most populated cluster corresponded to scores around 450. We therefore repeated the *holo* AM-6 MD simulation, this time using ligand and protein configurations associated with a LUDI score of 450 to create the initial frame. After 1.5 ns of equilibration, we let the simulation run another 1.35 ns and analyzed the AutoDock-predicted binding energy of the productive ensemble, as described above. As expected, the average AutoDock-predicted binding affinity was substantially lower when more representative protein and AM-6 ligand conformations were considered. In cases where the protein configurations associated with the best LUDI scores are rare, we therefore suggest using more representative protein and ligand configurations for subsequent *holo* simulations instead.

Generating Composite Compounds

The current scheme is amenable to the design of composite compounds in which optimum fragments targeting both the S1' and the S2'/S3' pockets are simultaneously added to the same ZBG (described as targeting “both” pockets in Table 11). In this study, a S1'-binding fragment and a S2'/S3'-binding fragment were selected to create a high-affinity composite compound. Fragment UD2a, which had the highest ensemble-average AutoDock binding affinity (Score^{*holo*}) of all novel hits, was selected as the S1'-binding fragment. As fragment U97 merged with the ZBG was the only S2'/S3'-binding ligand with a potential linker hydrogen pointed in the direction of the S1' pocket, U97 was selected as the S2'/S3' fragment. The composite compound

was created by simultaneously replacing the two hydrogen-atom linkers (Figure 28, inset figure) with the UD2a and U97 fragments, respectively, producing a chiral compound in the S configuration. The initial orientation of each fragment corresponded to the ensemble-maximum LUDI score for each. Two reasonable receptor configurations were envisioned for the ligand-protein complex: the optimal *apo* conformation corresponding to fragments U97 and UD2a, respectively (Figures 3A and 3B). To predict the binding affinity of our composite ligand, two sets of MD and AutoDock-rescoring calculations were performed, as described above, each starting from one of these initial protein structures. Figure 30C displays the final configuration of the MD simulation started from the optimal *apo* conformation for the UD2a fragment. A similar configuration was obtained when the optimal *apo* conformation for the U97 fragment was used.

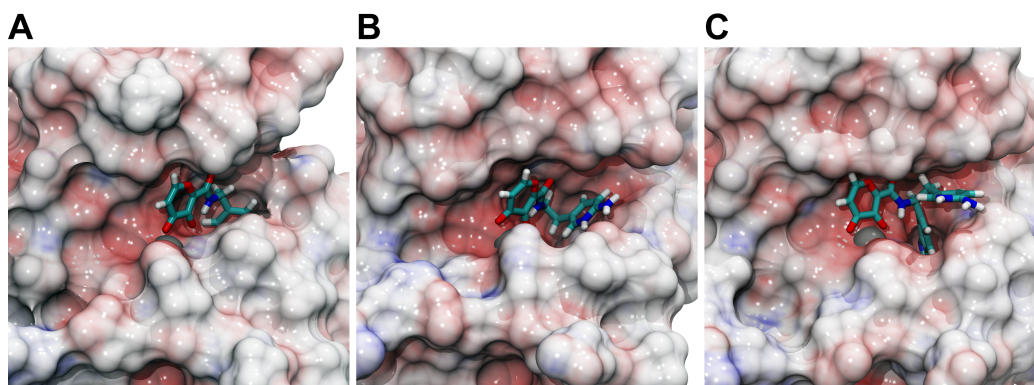


Figure 30: LUDI-docked fragments added to a ZBG. A) The ligand associated with fragment UD2a. B) The ligand associated with fragment U97. C) The equilibrated structure of the complex formed between the composite compound (UD2a + U97) and the last MMP-2 configuration of a *holo* MD simulation.

After equilibrating each of the two MD simulations of the composite compound (UD2a + U97), the ensemble-average scores ($\text{Score}^{\text{holo}}$) of both converged to nearly the same value. Accounting for induced-fit effects clearly has a substantial impact on the predicted binding energy by removing dependency on a single, often unrepresentative receptor conformation.

9.4 Conclusion

The rational design of MMP inhibitors is challenging due to the flexibility of the MMP active site. In this work, we combined computer docking with MD simulations in order to account for receptor flexibility, conformational selection mechanisms, and induced-fit effects. Our strategy identified fragments that bind the S1' pocket as well as fragments that bind elsewhere. Subsequently combining both types of fragments with an experimentally validated MMP ZBG group improved predicted potency. Our results indicate the importance of taking into account conformational selection and induced-fit effects when dealing with highly flexible proteins like the MMPs, as these factors can have a significant impact on the predicted binding energies of candidate compounds.

9.5 Acknowledgments

This work was supported in part by grants from the UCSD School of Medicine, NSF, NIH, the Center for Theoretical Biological Physics, the National

Biomedical Computation Resource, San Diego Supercomputing Center, and Accelrys, Inc. We thank Dr. Seth Cohen and his group for helpful discussions.

Chapter 9, in full, is a near reprint of the material as it appears in *Journal of Molecular Recognition* 2009, volume 23, issue 2, pages 173-182. The dissertation author was the primary contributor to this work. César Augusto F. de Oliveira also made significant contributions, and J. Andrew McCammon served as head adviser and principal investigator. The *Journal of Molecular Recognition* holds the copyright for this manuscript and has authorized its publication in this dissertation.

References

1. Trouiller, P.; Olliaro, P.; Torreele, E.; Orbinski, J.; Laing, R.; Ford, N. *Drug development for neglected diseases: a deficient market and a public-health policy failure* **2002**, Lancet, 359, 2188.
2. Remme, J. H.; Blas, E.; Chitsulo, L.; Desjeux, P. M.; Engers, H. D.; Kanyok, T. P.; Kengeya Kayondo, J. F.; Kioy, D. W.; Kumaraswami, V.; Lazdins, J. K.; Nunn, P. P.; Oduola, A.; Ridley, R. G.; Toure, Y. T.; Zicker, F.; Morel, C. M. *Strategic emphases for tropical diseases research: a TDR perspective* **2002**, Trends in parasitology, 18, 421-6.
3. del Rayo Camacho, M.; Phillipson, J. D.; Croft, S. L.; Rock, P.; Marshall, S. J.; Schiff, P. L., Jr. *In vitro activity of Triclisia patens and some bisbenzylisoquinoline alkaloids against Leishmania donovani and Trypanosoma brucei brucei* **2002**, Phytother Res, 16, 432-6.
4. Mehlert, A.; Zitzmann, N.; Richardson, J. M.; Treumann, A.; Ferguson, M. A. *The glycosylation of the variant surface glycoproteins and procyclic acidic repetitive proteins of Trypanosoma brucei* **1998**, Mol Biochem Parasitol, 91, 145-52.
5. Nolan, D. P.; Geuskens, M.; Pays, E. *N-linked glycans containing linear poly-N-acetyllactosamine as sorting signals in endocytosis in Trypanosoma brucei* **1999**, Curr Biol, 9, 1169-72.
6. Tetaud, E.; Barrett, M. P.; Bringaud, F.; Baltz, T. *Kinetoplastid glucose transporters* **1997**, Biochem J, 325 (Pt 3), 569-80.
7. Eienthal, R.; Game, S.; Holman, G. D. *Specificity and kinetics of hexose transport in Trypanosoma brucei* **1989**, Biochim Biophys Acta, 985, 81-9.
8. Liu, J. P. *Avian influenza--a pandemic waiting to happen?* **2006**, J Microbiol Immunol Infect, 39, 4-10.
9. Peiris, J. S.; de Jong, M. D.; Guan, Y. *Avian influenza virus (H5N1): a threat to human health* **2007**, Clin Microbiol Rev, 20, 243-67.
10. Moscona, A. *Neuraminidase inhibitors for influenza* **2005**, N Engl J Med, 353, 1363-73.

11. Senior, K. *Chagas disease: moving towards global elimination* **2007**, Lancet Infect Dis, 7, 572.
12. McKerrow, J. H.; McGrath, M. E.; Engel, J. C. *The cysteine protease of Trypanosoma cruzi as a model for antiparasite drug design* **1995**, Parasitol Today, 11, 279-82.
13. Harth, G.; Andrews, N.; Mills, A. A.; Engel, J. C.; Smith, R.; McKerrow, J. H. *Peptide-fluoromethyl ketones arrest intracellular replication and intercellular transmission of Trypanosoma cruzi* **1993**, Mol Biochem Parasitol, 58, 17-24.
14. Polticelli, F.; Zaini, G.; Bolli, A.; Antonini, G.; Gradoni, L.; Ascenzi, P. *Probing the cruzain S2 recognition subsite: a kinetic and binding energy calculation study* **2005**, Biochemistry, 44, 2781-9.
15. Tomas, A. M.; Miles, M. A.; Kelly, J. M. *Overexpression of cruzipain, the major cysteine proteinase of Trypanosoma cruzi, is associated with enhanced metacyclogenesis* **1997**, Eur J Biochem, 244, 596-603.
16. McKerrow, J. H.; Doyle, P. S.; Engel, J. C.; Podust, L. M.; Robertson, S. A.; Ferreira, R.; Saxton, T.; Arkin, M.; Kerr, I. D.; Brinen, L. S.; Craik, C. S. *Two approaches to discovering and developing new drugs for Chagas disease* **2009**, Mem Inst Oswaldo Cruz, 104, 263-9.
17. McKerrow, J. H.; Engel, J. C.; Caffrey, C. R. *Cysteine protease inhibitors as chemotherapy for parasitic infections* **1999**, Bioorg Med Chem, 7, 639-44.
18. McKerrow, J. H. *Development of cysteine protease inhibitors as chemotherapy for parasitic diseases: insights on safety, target validation, and mechanism of action* **1999**, Int J Parasitol, 29, 833-7.
19. Cazzulo, J. J. *Proteinases of Trypanosoma cruzi: potential targets for the chemotherapy of Chagas disease* **2002**, Curr Top Med Chem, 2, 1261-71.
20. Engel, J. C.; Doyle, P. S.; Hsieh, I.; McKerrow, J. H. *Cysteine protease inhibitors cure an experimental Trypanosoma cruzi infection* **1998**, J Exp Med, 188, 725-34.
21. Engel, J. C.; Doyle, P. S.; Palmer, J.; Hsieh, I.; Bainton, D. F.; McKerrow, J. H. *Cysteine protease inhibitors alter Golgi complex ultrastructure and function in Trypanosoma cruzi* **1998**, J Cell Sci, 111 (Pt 5), 597-606.

22. Chambers, A. F.; Matrisian, L. M. *Changing views of the role of matrix metalloproteinases in metastasis* **1997**, JNatlCancer Inst, 89, 1260-70.
23. Colasanti, M.; Salvati, L.; Venturini, G.; Ascenzi, P.; Gradoni, L. *Cysteine protease as a target for nitric oxide in parasitic organisms* **2001**, Trends Parasitol, 17, 575.
24. Ascenzi, P.; Salvati, L.; Bolognesi, M.; Colasanti, M.; Polticelli, F.; Venturini, G. *Inhibition of cysteine protease activity by NO-donors* **2001**, Curr Protein Pept Sci, 2, 137-53.
25. Barr, S. C.; Warner, K. L.; Kornreic, B. G.; Piscitelli, J.; Wolfe, A.; Benet, L.; McKerrow, J. H. *A cysteine protease inhibitor protects dogs from cardiac damage during infection by Trypanosoma cruzi* **2005**, Antimicrob Agents Chemother, 49, 5160-1.
26. Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. *Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function* **1998**, Journal of Computational Chemistry, 19, 1639-62.
27. Trott, O.; Olson, A. J. *AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading* **2009**, J Comput Chem, 31, 455-61.
28. Amaro, R. E.; Baron, R.; McCammon, J. A. *An improved relaxed complex scheme for receptor flexibility in computer-aided drug design* **2008**, JComputAided MolDes, 22, 693-705.
29. Lin, J. H.; Perryman, A. L.; Schames, J. R.; McCammon, J. A. *Computational drug design accommodating receptor flexibility: the relaxed complex scheme* **2002**, Journal of the American Chemical Society, 124, 5632-3.
30. Schames, J. R.; Henchman, R. H.; Siegel, J. S.; Sotriffer, C. A.; Ni, H.; McCammon, J. A. *Discovery of a novel binding trench in HIV integrase* **2004**, JMedChem, 47, 1879-81.
31. Amaro, R. E.; Schnaufer, A.; Interthal, H.; Hol, W.; Stuart, K. D.; McCammon, J. A. *Discovery of drug-like inhibitors of an essential RNA-editing ligase in Trypanosoma brucei* **2008**, Proceedings of the National Academy of Sciences, 105, 17278-83.

32. Kennedy, T. *Managing the drug discovery/development interface* **1997**, Drug DiscovToday, 2, 436-44.
33. WHO. African trypanosomiasis (sleeping sickness), Fact Sheet #259. WHO; 2006 [updated 2006 August 2006; cited 2009 Oct 12, 2009]; Available from: <http://www.who.int/mediacentre/factsheets/fs259/en/>.
34. Gehrig, S.; Efferth, T. *Development of drug resistance in Trypanosoma brucei rhodesiense and Trypanosoma brucei gambiense. Treatment of human African trypanosomiasis with natural products (Review)* **2008**, Int J Mol Med, 22, 411-9.
35. Maser, P.; Luscher, A.; Kaminsky, R. *Drug transport and drug resistance in African trypanosomes* **2003**, Drug Resist Updat, 6, 281-90.
36. Delespaux, V.; de Koning, H. P. *Drugs and drug resistance in African trypanosomiasis* **2007**, Drug Resist Updat, 10, 30.
37. Vansterkenburg, E. L.; Coppens, I.; Wilting, J.; Bos, O. J.; Fischer, M. J.; Janssen, L. H.; Opperdoes, F. R. *The uptake of the trypanocidal drug suramin in combination with low-density lipoproteins by Trypanosoma brucei and its possible mode of action* **1993**, Acta Trop, 54, 237-50.
38. Blum, J.; Nkunku, S.; Burri, C. *Clinical description of encephalopathic syndromes and risk factors for their occurrence and outcome during melarsoprol treatment of human African trypanosomiasis* **2001**, Trop Med Int Health, 6, 390-400.
39. Kayser, O.; Kiderlen, A. F.; Croft, S. L. *Natural products as antiparasitic drugs* **2003**, Parasitol Res, 90 Suppl 2, S55-62.
40. Iten, M.; Matovu, E.; Brun, R.; Kaminsky, R. *Innate lack of susceptibility of Ugandan Trypanosoma brucei rhodesiense to DL-alpha-difluoromethylornithine (DFMO)* **1995**, Trop Med Parasitol, 46, 190-4.
41. Priotto, G.; Kasparian, S.; Mutombo, W.; Ngouama, D.; Ghorashian, S.; Arnold, U.; Ghabri, S.; Baudin, E.; Buard, V.; Kazadi-Kyanza, S.; Ilunga, M.; Mutangala, W.; Pohlig, G.; Schmid, C.; Karunakara, U.; Torreele, E.; Kande, V. *Nifurtimox-eflornithine combination therapy for second-stage African Trypanosoma brucei gambiense trypanosomiasis: a multicentre, randomised, phase III, non-inferiority trial* **2009**, Lancet, 374, 56-64.

42. Cross, G. A. *Antigenic variation in trypanosomes: secrets surface slowly* **1996**, Bioessays, 18, 283-91.
43. Smith, D. H.; Pepin, J.; Stich, A. H. *Human African trypanosomiasis: an emerging public health crisis* **1998**, Br Med Bull, 54, 341-55.
44. Schneider, G.; Fechner, U. *Computer-based de novo design of drug-like molecules* **2005**, Nature Reviews Drug Discovery, 4, 649.
45. Shaw, M. P.; Bond, C. S.; Roper, J. R.; Gourley, D. G.; Ferguson, M. A.; Hunter, W. N. *High-resolution crystal structure of Trypanosoma brucei UDP-galactose 4'-epimerase: a potential target for structure-based development of novel trypanocides* **2003**, Mol Biochem Parasitol, 126, 173-80.
46. Holden, H. M.; Rayment, I.; Thoden, J. B. *Structure and function of enzymes of the Leloir pathway for galactose metabolism* **2003**, J Biol Chem, 278, 43885-8.
47. Roper, J. R.; Guthrie, M. L.; Milne, K. G.; Ferguson, M. A. *Galactose metabolism is essential for the African sleeping sickness parasite Trypanosoma brucei* **2002**, ProcNatlAcadSciUSA, 99, 5884-9.
48. Roper, J. R.; Guthrie, M. L.; Macrae, J. I.; Prescott, A. R.; Hallyburton, I.; Acosta-Serrano, A.; Ferguson, M. A. *The suppression of galactose metabolism in procyclic form Trypanosoma brucei causes cessation of cell growth and alters procyclin glycoprotein structure and copy number* **2005**, J Biol Chem, 280, 19728-36.
49. Dolinsky, T. J.; Czodrowski, P.; Li, H.; Nielsen, J. E.; Jensen, J. H.; Klebe, G.; Baker, N. A. *PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations* **2007**, Nucleic Acids Res, 35, W522-W5.
50. Walker, R. C.; de Souza, M. M.; Mercer, I. P.; Gould, I. R.; Klug, D. R. *Large and fast relaxations inside a protein: Calculation and measurement of reorganization energies in mcohol dehydrogenase* **202**, Journal of Physical Chemistry B, 106, 11658-65.
51. Pavelites, J. J.; Gao, J.; Bash, P. A.; Mackerell Jr., A. D. *A molecular mechanics force field for NAD⁺ NADH, and the pyrophosphate groups of nucleotides* **1998**, Journal of Computational Chemistry, 18, 221-39.

52. Ponder, J. W.; Case, D. A. *Force fields for protein simulations* **2003**, Adv Protein Chem, 66, 27-85.
53. Bayly, C. I.; Cieplak, P.; Cornell, W.; Kollman, P. A.
54. Case, D. A.; Cheatham, T. E., 3rd; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M., Jr.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. *The Amber biomolecular simulation programs* **2005**, J Comput Chem, 26, 1668-88.
55. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *Comparison of simple potential functions for simulating liquid water* **1983**, JChemPhys, 79, 926.
56. Feller, S. E.; Zhang, Y.; Pastor, R. W.; Brooks, B. R. *Constant pressure molecular dynamics simulation: The Langevin piston method* **1995**, JChemPhys, 103, 4613.
57. Darden, T.; York, D.; Pedersen, L. *Particle mesh Ewald: An $N, \tilde{O}(\log(N))$ method for Ewald sums in large systems* **1993**, JChemPhys, 98, 10089.
58. Kale, L., R. Skeel, M. Bhandarkar, R. Brunner, A. Gursoy, N. Krawetz, J. Phillips, A. Shinozaki, K. Varadarajan, and K. Schulten. *NAMD2: greater scalability for parallel molecular dynamics* **1999**, J Comput Phys, 151, 283-312.
59. Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K. *Scalable molecular dynamics with NAMD* **2005**, Journal of computational chemistry, 26, 1781-802.
60. Landon, M. R.; Amaro, R. E.; Baron, R.; Ngan, C. H.; Ozonoff, D.; McCammon, J. A.; Vajda, S. *Novel druggable hot spots in avian influenza neuraminidase H5N1 revealed by computational solvent mapping of a reduced and representative receptor ensemble* **2008**, Chem Biol Drug Des, 71, 106-16.
61. Christen, M.; Hunenberger, P. H.; Bakowies, D.; Baron, R.; Burgi, R.; Geerke, D. P.; Heinz, T. N.; Kastenholz, M. A.; Krautler, V.; Oostenbrink, C.; Peter, C.; Trzesniak, D.; van Gunsteren, W. F. *The GROMOS software for biomolecular simulation: GROMOS05* **2005**, J Comput Chem, 26, 1719-51.
62. Urbaniak, M. D.; Tabudravu, J. N.; Msaki, A.; Matera, K. M.; Brenk, R.; Jaspars, M.; Ferguson, M. A. *Identification of novel inhibitors of UDP-Glc 4'-epimerase, a validated drug target for african sleeping sickness* **2006**, Bioorg Med Chem Lett, 16, 5744-7.

63. Raz, B.; Iten, M.; Grether-Buhler, Y.; Kaminsky, R.; Brun, R. *The Alamar Blue assay to determine drug sensitivity of African trypanosomes (T.b. rhodesiense and T.b. gambiense) in vitro* **1997**, Acta Trop, 68, 139-47.
64. Dodson, G. G.; Lane, D. P.; Verma, C. S. *Molecular simulations of protein dynamics: new windows on mechanisms in biology* **2008**, EMBO Rep, 9, 144-50.
65. Okazaki, K.; Takada, S. *Dynamic energy landscape view of coupled binding and protein conformational change: induced-fit versus population-shift mechanisms* **2008**, Proc Natl Acad Sci U S A, 105, 11182-7.
66. Brahma, A.; Banerjee, N.; Bhattacharyya, D. *UDP-galactose 4-epimerase from Kluyveromyces fragilis--catalytic sites of the homodimeric enzyme are functional and regulated* **2009**, FEBS J, 276, 6725-40.
67. Hay, M.; Bhaduri, A. *UDP-Glucose 4-epimerase from Saccharomyces fragilis. Allosteric kinetics with UDP-glucose as substrate* **1975**, J Biol Chem, 250, 4373-5.
68. Feng, B. Y.; Simeonov, A.; Jadhav, A.; Babaoglu, K.; Inglese, J.; Shoichet, B. K.; Austin, C. P. *A high-throughput screen for aggregation-based inhibition in a large compound library* **2007**, J Med Chem, 50, 2385-90.
69. Surig, U.; Gaal, K.; Kostenis, E.; Trankle, C.; Mohr, K.; Holzgrabe, U. *Muscarinic allosteric modulators: atypical structure-activity-relationships in bispiridinium-type compounds* **2006**, Arch Pharm (Weinheim), 339, 207-12.
70. Langmead, C. J. *Screening for positive allosteric modulators: assessment of modulator concentration-response curves as a screening paradigm* **2007**, J Biomol Screen, 12, 668-76.
71. Pate, P. G.; Wolfson, J. S.; McHugh, G. L.; Pan, S. C.; Swartz, M. N. *Novobiocin antagonism of amastigotes of Trypanosoma cruzi growing in cell-free medium* **1986**, Antimicrob Agents Chemother, 29, 426-31.
72. Kerschmann, R. L.; Wolfson, J. S.; McHugh, G. L.; Dickersin, G. R.; Hooper, D. C.; Swartz, M. N. *Novobiocin-induced ultrastructural changes and antagonism of DNA synthesis in Trypanosoma cruzi amastigotes growing in cell-free medium* **1989**, J Protozool, 36, 14-20.
73. Dougherty, D. A. *Cation- π interactions in chemistry and biology: a new view of benzene, Phe, Tyr, and Trp* **1996**, Science, 271, 163-8.

74. Kerr, I. D.; Lee, J. H.; Farady, C. J.; Marion, R.; Rickert, M.; Sajid, M.; Pandey, K. C.; Caffrey, C. R.; Legac, J.; Hansell, E.; McKerrow, J. H.; Craik, C. S.; Rosenthal, P. J.; Brinen, L. S. *Vinyl sulfones as antiparasitic agents and a structural basis for drug design* **2009**, J Biol Chem, 284, 25697-703.
75. Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. *Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings* **2001**, Adv Drug Deliv Rev, 46, 3-26.
76. Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. *A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. A qualitative and quantitative characterization of known drug databases* **1999**, J Comb Chem, 1, 55-68.
77. Testa, B. *Prodrug research: futile or fertile?* **2004**, Biochem Pharmacol, 68, 2097-106.
78. Hotez, P. J.; Molyneux, D. H.; Fenwick, A.; Kumaresan, J.; Sachs, S. E.; Sachs, J. D.; Savioli, L. *Control of neglected tropical diseases* **2007**, N Engl J Med, 357, 1018-27.
79. Stich, A.; Abel, P. M.; Krishna, S. *Human African trypanosomiasis* **2002**, Bmj, 325, 203-6.
80. Legros, D.; Ollivier, G.; Gastellu-Etchegorry, M.; Paquet, C.; Burri, C.; Jannin, J.; Buscher, P. *Treatment of human African trypanosomiasis--present situation and needs for research and development* **2002**, Lancet Infect Dis, 2, 437-40.
81. WHO. The World Health Organisation: African trypanosomiasis, Fact sheet No. 259. Geneva: WHO publications; 2006.
82. Burri, C.; Keiser, J. *Pharmacokinetic investigations in patients from northern Angola refractory to melarsoprol treatment* **2001**, TropMedIntHealth, 6, 412.
83. McManus, M. T.; Shimamura, M.; Grams, J.; Hajduk, S. L. *Identification of candidate mitochondrial RNA editing ligases from Trypanosoma brucei* **2001**, Rna, 7, 167-75.
84. Swift, R. V.; Durrant, J.; Amaro, R. E.; McCammon, J. A. *Toward understanding the conformational dynamics of RNA ligation* **2009**, Biochemistry, 48, 709-19.

85. Irwin, J. J.; Shoichet, B. K. *ZINC--a free database of commercially available compounds for virtual screening* **2005**, JChemInfModel, 45, 177.
86. Deng, J.; Schnauffer, A.; Salavati, R.; Stuart, K. D.; Hol, W. G. *High resolution crystal structure of a key editosome enzyme from Trypanosoma brucei: RNA editing ligase I* **2004**, JMolBiol, 343, 601-13.
87. Amaro, R. E.; Swift, R. V.; McCammon, J. A. *Functional and Structural Insights Revealed by Molecular Dynamics Simulations of an Essential RNA Editing Ligase in Trypanosoma brucei* **2007**, PLoS Negl TropDis, 1, e68.
88. O'Donoghue, P.; Luthey-Schulten, Z. *On the evolution of structure in aminoacyl-tRNA synthetases* **2003**, Microbiol Mol Biol Rev, 67, 550-73.
89. Schnauffer, A.; Panigrahi, A. K.; Panicucci, B.; Igo, R. P., Jr.; Salavati, R.; Stuart, K. *An RNA Ligase Essential for RNA Editing and Survival of the Bloodstream Form of Trypanosoma brucei* **2001**, Science, 291, 2159-62.
90. Rusche, L. N.; Huang, C. E.; Piller, K. J.; Hemann, M.; Wirtz, E.; Sollner-Webb, B. *The two RNA ligases of the Trypanosoma brucei RNA editing complex: cloning the essential band IV gene and identifying the band V gene* **2001**, Mol Cell Biol, 21, 979-89.
91. Oostenbrink, B. C.; Pitera, J. W.; van Lipzig, M. M.; Meerman, J. H.; van Gunsteren, W. F. *Simulations of the estrogen receptor ligand-binding domain: affinity of natural ligands and xenoestrogens* **2000**, J Med Chem, 43, 4594-605.
92. Oostenbrink, C.; van Gunsteren, W. F. *Free energies of binding of polychlorinated biphenyls to the estrogen receptor from a single simulation* **2004**, Proteins, 54, 237-46.
93. Kim, J. T.; Hamilton, A. D.; Bailey, C. M.; Domaoal, R. A.; Wang, L.; Anderson, K. S.; Jorgensen, W. L. *FEP-guided selection of bicyclic heterocycles in lead optimization for non-nucleoside inhibitors of HIV-1 reverse transcriptase* **2006**, J Am Chem Soc, 128, 15372-3.
94. Bursulaya, B. D.; Totrov, M.; Abagyan, R.; Brooks, C. L., 3rd. *Comparative study of several algorithms for flexible ligand docking* **2003**, J Comput Aided Mol Des, 17, 755-63.
95. Ewing, T. J.; Makino, S.; Skillman, A. G.; Kuntz, I. D. *DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases*

- 2001**, JComputAided MolDes, 15, 411.
96. Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. *A fast flexible docking method using an incremental construction algorithm* **1996**, JMolBiol, 261, 470.
 97. Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. *Development and validation of a genetic algorithm for flexible docking* **1997**, JMolBiol, 267, 727.
 98. Barrett, M. P.; Boykin, D. W.; Brun, R.; Tidwell, R. R. *Human African trypanosomiasis: pharmacological re-engagement with a neglected disease* **2007**, BrJPharmacol, 152, 1155-71.
 99. Dawood, F. S.; Jain, S.; Finelli, L.; Shaw, M. W.; Lindstrom, S.; Garten, R. J.; Gubareva, L. V.; Xu, X.; Bridges, C. B.; Uyeki, T. M. *Emergence of a novel swine-origin influenza A (H1N1) virus in humans* **2009**, N Engl J Med, 360, 2605-15.
 100. Abdel-Ghafar, A. N.; Chotpitayasunondh, T.; Gao, Z.; Hayden, F. G.; Nguyen, D. H.; de Jong, M. D.; Naghdaliyev, A.; Peiris, J. S.; Shindo, N.; Soeroso, S.; Uyeki, T. M. *Update on avian influenza A (H5N1) virus infection in humans* **2008**, N Engl J Med, 358, 261-73.
 101. Oxford, J.; Balasingam, S.; Lambkin, R. *A new millennium conundrum: how to use a powerful class of influenza anti-neuraminidase drugs (NAIs) in the community* **2004**, Journal of Antimicrobial Chemotherapy, 53, 133-6.
 102. De Clercq, E.; Neyts, J. *Avian influenza A (H5N1) infection: targets and strategies for chemotherapeutic intervention* **2007**, Trends Pharmacol Sci, 28, 280-5.
 103. Kobasa, D.; Kodihalli, S.; Luo, M.; Castrucci, M. R.; Donatelli, I.; Suzuki, Y.; Suzuki, T.; Kawaoka, Y. *Amino acid residues contributing to the substrate specificity of the influenza A virus neuraminidase* **1999**, J Virol, 73, 6743-51.
 104. Beigel, J. H.; Farrar, J.; Han, A. M.; Hayden, F. G.; Hyer, R.; de Jong, M. D.; Lochindarat, S.; Nguyen, T. K.; Nguyen, T. H.; Tran, T. H.; Nicoll, A.; Touch, S.; Yuen, K. Y. *Avian influenza A (H5N1) infection in humans* **2005**, N Engl J Med, 353, 1374-85.
 105. de Jong, M. D.; Tran, T. T.; Truong, H. K.; Vo, M. H.; Smith, G. J.; Nguyen, V. C.; Bach, V. C.; Phan, T. Q.; Do, Q. H.; Guan, Y.; Peiris, J. S.; Tran, T. H.; Farrar, J. *Oseltamivir resistance during treatment of influenza A (H5N1)*

- infection* **2005**, N Engl J Med, 353, 2667-72.
106. De Clercq, E. *Antiviral agents active against influenza A viruses* **2006**, Nat Rev Drug Discov, 5, 1015-25.
 107. Kiso, M.; Mitamura, K.; Sakai-Tagawa, Y.; Shiraishi, K.; Kawakami, C.; Kimura, K.; Hayden, F. G.; Sugaya, N.; Kawaoka, Y. *Resistant influenza A viruses in children treated with oseltamivir: descriptive study* **2004**, The Lancet, 364, 759-65.
 108. Durrant, J. D.; Amaro, R. E.; McCammon, J. A. *AutoGrow: A Novel Algorithm for Protein Inhibitor Design* **2009**, Chemical Biology & Drug Design, 73, 168-78.
 109. Cheng, L. S.; Amaro, R. E.; Xu, D.; Li, W. W.; Arzberger, P. W.; McCammon, J. A. *Ensemble-based virtual screening reveals potential novel antiviral compounds for avian influenza neuraminidase* **2008**, J Med Chem, 51, 3878-94.
 110. Daura, X.; Gademann, K.; Jaun, B.; Seebach, D.; van Gunsteren, W. F.; Mark, A. E. *Peptide Folding: When Simulation Meets Experiment* **1999**, Angew Chem Int Ed, 38, 236-40.
 111. Russell, R. J.; Haire, L. F.; Stevens, D. J.; Collins, P. J.; Lin, Y. P.; Blackburn, G. M.; Hay, A. J.; Gamblin, S. J.; Skehel, J. J. *The structure of H5N1 avian influenza neuraminidase suggests new opportunities for drug design* **2006**, Nature, 443, 45-9.
 112. Wishart, D. S.; Knox, C.; Guo, A. C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z.; Woolsey, J. *DrugBank: a comprehensive resource for in silico drug discovery and exploration* **2006**, Nucleic Acids Res, 34, D668-72.
 113. Varghese, J. N.; Laver, W. G.; Colman, P. M. *Structure of the influenza virus glycoprotein antigen neuraminidase at 2.9 Å resolution* **1983**, Nature, 303, 35-40.
 114. Baker, A. T.; Varghese, J. N.; Laver, W. G.; Air, G. M.; Colman, P. M. *Three-dimensional structure of neuraminidase of subtype N9 from an avian influenza virus* **1987**, Proteins, 2, 111-7.
 115. Amaro, R. E.; Minh, D. D.; Cheng, L. S.; Lindstrom, W. M., Jr.; Olson, A. J.; Lin, J. H.; Li, W. W.; McCammon, J. A. *Remarkable loop flexibility in avian influenza N1 and its implications for antiviral drug design* **2007**, J Am Chem

Soc, 129, 7764-5.

116. Rees, D. C.; Congreve, M.; Murray, C. W.; Carr, R. *Fragment-based lead discovery* **2004**, Nat Rev Drug Discov, 3, 660-72.
117. Bohm, H.-J. *The computer program LUDI: A new method for the de novo design of enzyme inhibitors* **1992**, JComputAided MolDes, 6, 61.
118. Bohm, H. J. *A novel computational tool for automated structure-based drug design* **1993**, JMolRecognit, 6, 131.
119. Bohm, H. J. *On the use of LUDI to search the Fine Chemicals Directory for ligands of proteins of known three-dimensional structure* **1994**, JComputAided MolDes, 8, 623-32.
120. Malaisree, M.; Rungrotmongkol, T.; Decha, P.; Intharathep, P.; Aruksakunwong, O.; Hannongbua, S. *Understanding of known drug-target interactions in the catalytic pocket of neuraminidase subtype N1* **2008**, Proteins, 71, 1908-18.
121. Colman, P. M. *Influenza virus neuraminidase: structure, antibodies, and inhibitors* **1994**, Protein Sci, 3, 1687-96.
122. Zhang, X. J.; Baase, W. A.; Matthews, B. W. *Multiple alanine replacements within alpha-helix 126-134 of T4 lysozyme have independent, additive effects on both structure and stability* **1992**, Protein Sci, 1, 761-76.
123. Xu, X.; Zhu, X.; Dwek, R. A.; Stevens, J.; Wilson, I. A. *Structural characterization of the 1918 influenza virus H1N1 neuraminidase* **2008**, J Virol, 82, 10493-501.
124. Moscona, A. *Oseltamivir resistance--disabling our influenza defenses* **2005**, N Engl J Med, 353, 2633-6.
125. WHO. *Control of Chagas disease* **2002**, World Health Organ Tech Rep Ser, 905, i-vi, 1-109, back cover.
126. McKerrow, J. H.; Caffrey, C.; Kelly, B.; Loke, P.; Sajid, M. *Proteases in parasitic diseases* **2006**, Annu Rev Pathol, 1, 497-536.
127. Du, X.; Guo, C.; Hansell, E.; Doyle, P. S.; Caffrey, C. R.; Holler, T. P.; McKerrow, J. H.; Cohen, F. E. *Synthesis and structure-activity relationship study of potent trypanocidal thio semicarbazone inhibitors of the trypanosomal*

- cysteine protease cruzain* **2002**, J Med Chem, 45, 2695-707.
128. Rodriques Coura, J.; de Castro, S. L. *A critical review on Chagas disease chemotherapy* **2002**, Mem Inst Oswaldo Cruz, 97, 3-24.
 129. Bonney, K. M.; Engman, D. M. *Chagas heart disease pathogenesis: one mechanism or many?* **2008**, Curr Mol Med, 8, 510-8.
 130. de Castro, S. L. *The challenge of Chagas' disease chemotherapy: an update of drugs assayed against Trypanosoma cruzi* **1993**, Acta Trop, 53, 83-98.
 131. Wilkinson, S. R.; Taylor, M. C.; Horn, D.; Kelly, J. M.; Cheeseman, I. *A mechanism for cross-resistance to nifurtimox and benznidazole in trypanosomes* **2008**, Proc Natl Acad Sci U S A, 105, 5022-7.
 132. Lauria-Pires, L.; Braga, M. S.; Vexenat, A. C.; Nitz, N.; Simoes-Barbosa, A.; Tinoco, D. L.; Teixeira, A. R. *Progressive chronic Chagas heart disease ten years after treatment with anti-Trypanosoma cruzi nitroderivatives* **2000**, Am J Trop Med Hyg, 63, 111-8.
 133. Huang, L.; Brinen, L. S.; Ellman, J. A. *Crystal structures of reversible ketone-based inhibitors of the cysteine protease cruzain* **2003**, Bioorg Med Chem, 11, 21-9.
 134. Soares, M. J. *The reservosome of Trypanosoma cruzi epimastigotes: an organelle of the endocytic pathway with a role on metacyclogenesis* **1999**, Mem Inst Oswaldo Cruz, 94 Suppl 1, 139-41.
 135. Dolinsky, T. J.; Nielsen, J. E.; McCammon, J. A.; Baker, N. A. *PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations* **2004**, Nucleic Acids Res, 32, W665-W7.
 136. Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *Development and testing of a general amber force field* **2004**, J Comput Chem, 25, 1157-74.
 137. Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Comparison of multiple Amber force fields and development of improved protein backbone parameters* **2006**, Proteins, 65, 712-25.
 138. Duschak, V. G.; Couto, A. S. *Cruzipain, the major cysteine protease of Trypanosoma cruzi: a sulfated glycoprotein antigen as relevant candidate for vaccine development and drug target. A review* **2009**, Curr Med Chem, 16,

3174-202.

139. Gehlhaar, D. K.; Verkhivker, G. M.; Rejto, P. A.; Sherman, C. J.; Fogel, D. B.; Fogel, L. J.; Freer, S. T. *Molecular recognition of the inhibitor AG-1343 by HIV-1 protease: conformationally flexible docking by evolutionary programming* **1995**, Chem Biol, 2, 317-24.
140. Bairoch, A.; Apweiler, R.; Wu, C. H.; Barker, W. C.; Boeckmann, B.; Ferro, S.; Gasteiger, E.; Huang, H.; Lopez, R.; Magrane, M.; Martin, M. J.; Natale, D. A.; O'Donovan, C.; Redaschi, N.; Yeh, L.-S. L. *The Universal Protein Resource (UniProt)* **2005**, Nucl Acids Res, 33, D154-9.
141. Rawlings, N. D.; Morton, F. R.; Kok, C. Y.; Kong, J.; Barrett, A. J. *MEROPS: the peptidase database* **2008**, Nucleic Acids Res, 36, D320-5.
142. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *The Protein Data Bank* **2000**, Nucl Acids Res, 28, 235-42.
143. Larkin, M. A.; Blackshields, G.; Brown, N. P.; Chenna, R.; McGettigan, P. A.; McWilliam, H.; Valentin, F.; Wallace, I. M.; Wilm, A.; Lopez, R.; Thompson, J. D.; Gibson, T. J.; Higgins, D. G. *Clustal W and Clustal X version 2.0* **2007**, Bioinformatics, 23, 2947-8.
144. Roberts, E.; Eargle, J.; Wright, D.; Luthey-Schulten, Z. *MultiSeq: unifying sequence and structure data for evolutionary analysis* **2006**, BMC Bioinformatics, 7, 382.
145. Humphrey, W.; Dalke, A.; Schulten, K. *VMD: visual molecular dynamics* **1996**, J Mol Graph, 14, 33-8.
146. Russell, R. B.; Barton, G. J. *Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels* **1992**, Proteins, 14, 309-23.
147. Brinen, L. S.; Hansell, E.; Cheng, J.; Roush, W. R.; McKerrow, J. H.; Fletterick, R. J. *A target within the target: probing cruzain's P1' site to define structural determinants for the Chagas' disease protease* **2000**, Structure, 8, 831-40.
148. Turk, D.; Guncar, G.; Podobnik, M.; Turk, B. *Revised definition of substrate binding sites of papain-like cysteine proteases* **1998**, Biol Chem, 379, 137-47.

149. Douc-Rasy, S.; Riou, J. F.; Ahomadegbe, J. C.; Riou, G. *ATP-independent DNA topoisomerase II as potential drug target in trypanosomes* **1988**, *Biol Cell*, 64, 145-56.
150. Gillmor, S. A.; Craik, C. S.; Fletterick, R. J. *Structural determinants of specificity in the cysteine protease cruzain* **1997**, *Protein Sci*, 6, 1603-11.
151. Allende, L. M.; Garcia-Perez, M. A.; Moreno, A.; Corell, A.; Carasol, M.; Martinez-Canut, P.; Arnaiz-Villena, A. *Cathepsin C gene: First compound heterozygous patient with Papillon-Lefevre syndrome and a novel symptomless mutation* **2001**, *Hum Mutat*, 17, 152-3.
152. Yang, Y.; Bai, X.; Liu, H.; Li, L.; Cao, C.; Ge, L. *Novel mutations of cathepsin C gene in two Chinese patients with Papillon-Lefevre syndrome* **2007**, *J Dent Res*, 86, 735-8.
153. Hart, T. C.; Hart, P. S.; Michalec, M. D.; Zhang, Y.; Firatli, E.; Van Dyke, T. E.; Stabholz, A.; Zlotogorski, A.; Shapira, L.; Soskolne, W. A. *Haim-Munk syndrome and Papillon-Lefevre syndrome are allelic mutations in cathepsin C* **2000**, *J Med Genet*, 37, 88-94.
154. Toomes, C.; James, J.; Wood, A. J.; Wu, C. L.; McCormick, D.; Lench, N.; Hewitt, C.; Moynihan, L.; Roberts, E.; Woods, C. G.; Markham, A.; Wong, M.; Widmer, R.; Ghaffar, K. A.; Pemberton, M.; Hussein, I. R.; Temtamy, S. A.; Davies, R.; Read, A. P.; Sloan, P.; Dixon, M. J.; Thakker, N. S. *Loss-of-function mutations in the cathepsin C gene result in periodontal disease and palmoplantar keratosis* **1999**, *Nat Genet*, 23, 421-4.
155. Stuart, K.; Allen, T. E.; Heidmann, S.; Seiwert, S. D. *RNA editing in kinetoplastid protozoa* **1997**, *MicrobiolMolBiolRev*, 61, 105-20.
156. Panigrahi, A. K.; Ernst, N. L.; Domingo, G. J.; Fleck, M.; Salavati, R.; Stuart, K. D. *Compositionally and functionally distinct editosomes in Trypanosoma brucei* **2006**, *Rna*, 12, 1038-49.
157. Blum, B.; Simpson, L. *Formation of guide RNA/messenger RNA chimeric molecules in vitro, the initial step of RNA editing, is dependent on an anchor sequence* **1992**, *ProcNatlAcadSciUSA*, 89, 11944-8.
158. Seiwert, S. D.; Heidmann, S.; Stuart, K. *Direct visualization of uridylyate deletion in vitro suggests a mechanism for kinetoplastid RNA editing* **1996**, *Cell*, 84, 831-41.

159. Simpson, L.; Sbicego, S.; Aphasizhev, R. *Uridine insertion/deletion RNA editing in trypanosome mitochondria: a complex business* **2003**, *Rna*, 9, 265-76.
160. Stuart, K. D.; Schnauffer, A.; Ernst, N. L.; Panigrahi, A. K. *Complex management: RNA editing in trypanosomes* **2005**, *Trends BiochemSci*, 30, 97-105.
161. Panigrahi, A. K.; Gygi, S. P.; Ernst, N. L.; Igo, R. P., Jr.; Palazzo, S. S.; Schnauffer, A.; Weston, D. S.; Carmean, N.; Salavati, R.; Aebersold, R.; Stuart, K. D. *Association of Two Novel Proteins, TbMP52 and TbMP48, with the Trypanosoma brucei RNA Editing Complex* **2001**, *Mol Cell Biol*, 21, 380.
162. Shuman, S.; Lima, C. D. *The polynucleotide ligase and RNA capping enzyme superfamily of covalent nucleotidyltransferases* **2004**, *Curr Opin Struct Biol*, 14, 757-64.
163. Hakansson, K.; Doherty, A. J.; Shuman, S.; Wigley, D. B. *X-ray crystallography reveals a large conformational change during guanyl transfer by mRNA capping enzymes* **1997**, *Cell*, 89, 545-53.
164. Cherepanov, A. V.; de Vries, S. *Kinetic mechanism of the Mg²⁺-dependent nucleotidyl transfer catalyzed by T4 DNA and RNA ligases* **2002**, *J Biol Chem*, 277, 1695-704.
165. Nandakumar, J.; Shuman, S.; Lima, C. D. *RNA ligase structures reveal the basis for RNA specificity and conformational changes that drive ligation forward* **2006**, *Cell*, 127, 71.
166. Pascal, J. M. *DNA and RNA ligases: structural variations and shared mechanisms* **2008**, *Curr Opin Struct Biol*, 18, 96-105.
167. Norberg, J.; Nilsson, L. *Molecular dynamics applied to nucleic acids* **2002**, *Acc Chem Res*, 35, 465-72.
168. Mackerell, A. D., Jr.; Nilsson, L. *Molecular dynamics simulations of nucleic acid-protein complexes* **2008**, *Curr Opin Struct Biol*, 18, 194-9.
169. Case, D. A.; Darden, T. A.; Cheatham Iii, T. E.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Merz, K. M.; Pearlman, D. A.; Crowley, M.; Walker, R. C.; Zhang, W.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Wong, K. F.; Paesani, F.; Wu, X.; Brozell, S.; Tsui, V.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Beroza, P.; Mathews, D. H.; Schafmeister,

- C.; Ross, W. S.; Kollman, P. A. *Amber 9* **2006**,
170. El Omari, K.; Ren, J.; Bird, L. E.; Bona, M. K.; Klarmann, G.; LeGrice, S. F.; Stammers, D. K. *Molecular architecture and ligand recognition determinants for T4 RNA ligase* **2006**, J Biol Chem, 281, 1573-9.
 171. Vriend, G. *WHAT IF: a molecular modeling and drug design program* **1990**, J Mol Graph, 8, 52-6, 29.
 172. Meagher, K. L.; Redman, L. T.; Carlson, H. A. *Development of polyphosphate parameters for use with the AMBER force field* **2003**, J Comput Chem, 24, 1016-25.
 173. Rao, S. T.; Westhof, E.; Sundaralingam, M. *Exact method for the calculation of pseudorotation parameters P_{tm} and their errors. A comparison of the Altona-Sundaralingam and Cremer-Pople treatment of puckering of five-membered rings.* **1981**, Acta Crystallogr, A37, 421-5.
 174. Arora, K.; Schlick, T. *Deoxyadenosine sugar puckering pathway simulated by the stochastic difference equation algorithm* **2003**, Chemical Physics Letters, 378, 1-8.
 175. Amadei, A.; Linssen, A. B.; Berendsen, H. J. *Essential dynamics of proteins* **1993**, Proteins, 17, 412-25.
 176. Garcia, A. E. *Large-amplitude nonlinear motions in proteins* **1992**, Phys Rev Lett, 68, 2696-9.
 177. Yin, S.; Ho, C. K.; Shuman, S. *Structure-function analysis of T4 RNA ligase 2* **2003**, J Biol Chem, 278, 17601-8.
 178. Nandakumar, J.; Shuman, S. *How an RNA ligase discriminates RNA versus DNA damage* **2004**, Mol Cell, 16, 211-21.
 179. Palazzo, S. S.; Panigrahi, A. K.; Igo, R. P.; Salavati, R.; Stuart, K. *Kinetoplastid RNA editing ligases: complex association, characterization, and substrate requirements* **2003**, Mol Biochem Parasitol, 127, 161-7.
 180. Allen, K. N.; Dunaway-Mariano, D. *Phosphoryl group transfer: evolution of a catalytic scaffold* **2004**, Trends Biochem Sci, 29, 495-503.
 181. Fersht, A. *Structure and Mechanism in Protein Science*. New York: W. H. Freeman and Company; 1999.

182. Bairoch, A. *The ENZYME database in 2000* **2000**, Nucleic Acids Res, 28, 304-5.
183. Adams, C. P.; Brantner, V. V. *Estimating the cost of new drug development: is it really 802 million dollars?* **2006**, Health Aff (Millwood), 25, 420-8.
184. Keiser, M. J.; Roth, B. L.; Armbruster, B. N.; Ernsberger, P.; Irwin, J. J.; Shoichet, B. K. *Relating protein pharmacology by ligand chemistry* **2007**, Nat Biotechnol, 25, 197-206.
185. Croft, S. L. *Public-private partnership: from there to here* **2005**, Trans R Soc Trop Med Hyg, 99 Suppl 1, S9-14.
186. Schreiber, S. L. *Small molecules: the missing link in the central dogma* **2005**, Nat Chem Biol, 1, 64-6.
187. Paolini, G. V.; Shapland, R. H.; van Hoorn, W. P.; Mason, J. S.; Hopkins, A. L. *Global mapping of pharmacological space* **2006**, Nat Biotechnol, 24, 805-15.
188. Vieth, M.; Higgs, R. E.; Robertson, D. H.; Shapiro, M.; Gragg, E. A.; Hemmerle, H. *Kinomics-structural biology and chemogenomics of kinase inhibitors and targets* **2004**, Biochim Biophys Acta, 1697, 243-57.
189. Izrailev, S.; Farnum, M. A. *Enzyme classification by ligand binding* **2004**, Proteins, 57, 711-24.
190. Coleman, R. G.; Sharp, K. A. *Travel depth, a new shape descriptor for macromolecules: application to ligand binding* **2006**, J Mol Biol, 362, 441-58.
191. Nayal, M.; Honig, B. *On the nature of cavities on protein surfaces: application to the identification of drug-binding sites* **2006**, Proteins, 63, 892-906.
192. Coleman, R. G.; Burr, M. A.; Souvaine, D. L.; Cheng, A. C. *An intuitive approach to measuring protein surface curvature* **2005**, Proteins, 61, 1068-74.
193. Agarwal, P. K.; Edelsbrunner, H.; Harer, J.; Wang, Y. *Extreme elevation on a 2-manifold.* **2004**, Symp Comp Geo, 20, 357-65.
194. Hendrix, D. K.; Kuntz, I. D. *Surface solid angle-based site points for molecular docking* **1998**, Pac Symp Biocomput, 317-26.
195. Liang, J.; Edelsbrunner, H.; Woodward, C. *Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand*

- design* **1998**, Protein Sci, 7, 1884-97.
196. Norel, R.; Wolfson, H. J.; Nussinov, R. *Small molecule recognition: solid angles surface representation and molecular shape complementarity* **1999**, Comb Chem High Throughput Screen, 2, 223-37.
 197. Watson, J. D.; Laskowski, R. A.; Thornton, J. M. *Predicting protein function from sequence and structural data* **2005**, Curr Opin Struct Biol, 15, 275-84.
 198. Xie, L.; Bourne, P. *A robust and efficient algorithm for the shape description of protein structures and its application in predicting ligand binding sites* **2007**, BMC Bioinformatics, 8, S9.
 199. Xie, L.; Bourne, P. E. *Detecting evolutionary relationships across existing fold space, using sequence order-independent profile-profile alignments* **2008**, Proc Natl Acad Sci U S A, 105, 5441-6.
 200. Xie, L.; Wang, J.; Bourne, P. E. *In silico elucidation of the molecular mechanism defining the adverse effect of selective estrogen receptor modulators* **2007**, PLoS Comput Biol, 3, e217.
 201. Miinalainen, I. J.; Chen, Z. J.; Torkko, J. M.; Pirila, P. L.; Sormunen, R. T.; Bergmann, U.; Qin, Y. M.; Hiltunen, J. K. *Characterization of 2-enoyl thioester reductase from mammals. An ortholog of YBR026p/MRF1'p of the yeast mitochondrial fatty acid synthesis type II* **2003**, J Biol Chem, 278, 20154-61.
 202. Rake, J.; Visser, G.; Smit, G. Disorders of Carbohydrate and Glycogen Metabolism. Physician's Guide to the Treatment and Follow-Up of Metabolic Diseases; 2006. p. 161-80.
 203. Pascal, J. M.; O'Brien, P. J.; Tomkinson, A. E.; Ellenberger, T. *Human DNA ligase I completely encircles and partially unwinds nicked DNA* **2004**, Nature, 432, 473-8.
 204. Willis, A. E.; Lindahl, T. *DNA ligase I deficiency in Bloom's syndrome* **1987**, Nature, 325, 355-7.
 205. Harrison, C.; Ketchen, A. M.; Redhead, N. J.; O'Sullivan, M. J.; Melton, D. W. *Replication failure, genome instability, and increased cancer susceptibility in mice with a point mutation in the DNA ligase I gene* **2002**, Cancer Res, 62, 4065-74.

206. Ye, Y.; Godzik, A. *Flexible structure alignment by chaining aligned fragment pairs allowing twists* **2003**, Bioinformatics, 19 Suppl 2, ii246.
207. Allard, S. T.; Giraud, M. F.; Whitfield, C.; Graninger, M.; Messner, P.; Naismith, J. H. *The crystal structure of dTDP-D-Glucose 4,6-dehydratase (RmlB) from Salmonella enterica serovar Typhimurium, the second enzyme in the dTDP-l-rhamnose pathway* **2001**, J Mol Biol, 307, 283-95.
208. Gajiwala, K. S.; Pinko, C. *Structural rearrangement accompanying NAD⁺ synthesis within a bacterial DNA ligase crystal* **2004**, Structure, 12, 1449-59.
209. Xie, L.; Bourne, P. E. *A unified statistical model to support local sequence order independent similarity searching for ligand-binding sites and its application to genome-based drug discovery* **2009**, Bioinformatics, 25, i305-12.
210. Xie, L.; Li, J.; Bourne, P. E. *Drug discovery using chemical systems biology: identification of the protein-ligand binding network to explain the side effects of CETP inhibitors* **2009**, PLoS Comput Biol, 5, e1000387.
211. Porter, C. T.; Bartlett, G. J.; Thornton, J. M. *The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data* **2004**, Nucleic Acids Res, 32, D129-33.
212. Lau, A. Y.; Wyatt, M. D.; Glassner, B. J.; Samson, L. D.; Ellenberger, T. *Molecular basis for discriminating between normal and damaged bases by the human alkyladenine glycosylase, AAG* **2000**, Proc Natl Acad Sci U S A, 97, 13573-8.
213. Silverman, B. W. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall/CRC; 1986.
214. Jorgensen, W. L. *The many roles of computation in drug discovery* **2004**, Science, 303, 1813-8.
215. von Lilienfeld, O. A.; Lins, R. D.; Rothlisberger, U. *Variational particle number approach for rational compound design* **2005**, Phys Rev Lett, 95, 153002.
216. Wang, M.; Hu, X.; Beratan, D. N.; Yang, W. *Designing molecules by optimizing potentials* **2006**, J Am Chem Soc, 128, 3228-32.

217. Gillet, V. A.; Johnson, A. P.; Mata, P.; Sike, S. *Automated structure design in 3D* **1990**, Tetrahedron, 3, 681.
218. Moon, J. B.; Howe, W. J. *Computer design of bioactive molecules: A method for receptor-based de novo ligand design* **1991**, Proteins: Structure, Function, and Genetics, 11, 314.
219. Gillet, V.; Johnson, A. P.; Mata, P.; Sike, S.; Williams, P. *SPROUT: a program for structure generation* **1993**, JComputAided MolDes, 7, 127.
220. Rotstein, S. H.; Murcko, M. A. *GroupBuild: a fragment-based method for de novo drug design* **1993**, JMedChem, 36, 1700.
221. Bohacek, R. S.; McMartin, C. *Multiple Highly Diverse Structures Complementary to Enzyme Binding-Sites - Results of Extensive Application of a de-Novo Design Method Incorporating Combinatorial Growth* **1994**, Journal of the American Chemical Society, 116, 5560.
222. Clark, D. E.; Frenkel, D.; Levy, S. A.; Li, J.; Murray, C. W.; Robson, B.; Waszkowycz, B.; Westhead, D. R. *PRO_LIGAND: An approach to de novo molecular design. 1. Application to the design of organic molecules* **1995**, JComputAided MolDes, 9, 13.
223. Westhead, D. R.; Clark, D. E.; Frenkel, D.; Li, J.; Murray, C. W.; Robson, B.; Waszkowycz, B. *PRO_LIGAND: An approach to de novo molecular design. 3. A genetic algorithm for structure refinement* **1995**, JComputAided MolDes, 9, 139.
224. DeWitte, R. S.; Shakhnovich, E. I. *SMoG: de Novo Design Method Based on Simple, Fast, and Accurate Free Energy Estimates. 1. Methodology and Supporting Evidence* **1996**, Journal of the American Chemical Society, 118, 11733.
225. Wang, J.; Cieplak, P.; Kollman, P. A. *How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules?* **2000**, Journal of Computational Chemistry, 21, 1049-74.
226. Li, C.; Xu, L.; Wolan, D. W.; Wilson, I. A.; Olson, A. J. *Virtual screening of human 5-aminoimidazole-4-carboxamide ribonucleotide transformylase against the NCI diversity set by use of AutoDock to identify novel nonfolate inhibitors* **2004**, J Med Chem, 47, 6681-90.

227. Holland, J. H. *Adaptation in Natural and Artificial Systems*. Ann Arbor MI: University of Michigan Press; 1975.
228. Sousa, S. F.; Fernandes, P. A.; Ramos, M. J. *Protein-ligand docking: current status and future challenges* **2006**, *Proteins*, 65, 15-26.
229. Schaftenaar, G.; Noordik, J. H. *Molden: a pre- and post-processing program for molecular and electronic structures* **2000**, *J Comput Aided Mol Des*, 14, 123-34.
230. Schuttelkopf, A. W.; van Aalten, D. M. *PRODRG: a tool for high-throughput crystallography of protein-ligand complexes* **2004**, *Acta Crystallogr D Biol Crystallogr*, 60, 1355-63.
231. Sanner, M. F. *A component-based software environment for visualizing large macromolecular assemblies* **2005**, *Structure*, 13, 447-62.
232. Gasteiger, J.; Marsili, M. *Iterative partial equalization of orbital electronegativity--a rapid access to atomic charges* **1980**, *Tetrahedron*, 36, 3219-28.
233. Cheong, C. G.; Wolan, D. W.; Greasley, S. E.; Horton, P. A.; Beardsley, G. P.; Wilson, I. A. *Crystal structures of human bifunctional enzyme aminoimidazole-4-carboxamide ribonucleotide transformylase/IMP cyclohydrolase in complex with potent sulfonyl-containing antifolates* **2004**, *J Biol Chem*, 279, 18034-45.
234. Glen, R. C.; Payne, A. W. *A genetic algorithm for the automated generation of molecules within constraints* **1995**, *J Comput Aided Mol Des*, 9, 181-202.
235. Tschinke, V.; Cohen, N. C. *The NEWLEAD program: a new method for the design of candidate structures from pharmacophoric hypotheses* **1993**, *J Med Chem*, 36, 3863-70.
236. Wang, R.; Gao, Y.; Lai, L. *LigBuilder: A Multi-Purpose Program for Structure-Based Drug Design* **2000**, *Journal of Molecular Modeling*, 6, 498.
237. Pegg, S. C.; Haresco, J. J.; Kuntz, I. D. *A genetic algorithm for structure-based de novo design* **2001**, *J Comput Aided Mol Des*, 15, 911-33.
238. Vinkers, H. M.; de Jonge, M. R.; Daeyaert, F. F.; Heeres, J.; Koymans, L. M.; van Lenthe, J. H.; Lewi, P. J.; Timmerman, H.; Van Aken, K.; Janssen, P. A. *SYNOPSIS: SYNthesize and OPTimize System in Silico* **2003**, *J Med Chem*, 46,

2765-73.

239. Khodade, P.; Prabhu, R.; Chandra, N.; Raha, S.; Govindarajan, R. *Parallel implementation of AutoDock* **2007**, Journal of Applied Crystallography, 40, 598-9.
240. Moy, F. J.; Pisano, M. R.; Chanda, P. K.; Urbano, C.; Killar, L. M.; Sung, M. L.; Powers, R. *Assignments, secondary structure and dynamics of the inhibitor-free catalytic fragment of human fibroblast collagenase* **1997**, J Biomol NMR, 10, 9-19.
241. Moy, F. J.; Chanda, P. K.; Cosmi, S.; Pisano, M. R.; Urbano, C.; Wilhelm, J.; Powers, R. *High-resolution solution structure of the inhibitor-free catalytic fragment of human fibroblast collagenase determined by multidimensional NMR* **1998**, Biochemistry, 37, 1495-504.
242. Lovejoy, B.; Welch, A. R.; Carr, S.; Luong, C.; Broka, C.; Hendricks, R. T.; Campbell, J. A.; Walker, K. A.; Martin, R.; Van Wart, H.; Browner, M. F. *Crystal structures of MMP-1 and -13 reveal the structural basis for selectivity of collagenase inhibitors* **1999**, Nat Struct Biol, 6, 217-21.
243. Moy, F. J.; Chanda, P. K.; Chen, J. M.; Cosmi, S.; Edris, W.; Levin, J. I.; Powers, R. *High-resolution solution structure of the catalytic fragment of human collagenase-3 (MMP-13) complexed with a hydroxamic acid inhibitor* **2000**, J Mol Biol, 302, 671-89.
244. Zhang, X.; Gonnella, N. C.; Koehn, J.; Pathak, N.; Ganu, V.; Melton, R.; Parker, D.; Hu, S. I.; Nam, K. Y. *Solution structure of the catalytic domain of human collagenase-3 (MMP-13) complexed to a potent non-peptidic sulfonamide inhibitor: binding comparison with stromelysin-1 and collagenase-1* **2000**, J Mol Biol, 301, 513-24.
245. Matter, H.; Schudok, M. *Recent advances in the design of matrix metalloprotease inhibitors* **2004**, Curr Opin Drug Discov Devel, 7, 513-35.
246. Chang, C.; Werb, Z. *The many faces of metalloproteases: cell growth, invasion, angiogenesis and metastasis* **2001**, Trends Cell Biol, 11, S37-S43.
247. Coussens, L. M.; Fingleton, B.; Matrisian, L. M. *Matrix metalloproteinase inhibitors and cancer: trials and tribulations* **2002**, Science, 295, 2387-92.
248. Kahari, V. M.; Saarialho-Kere, U. *Matrix metalloproteinases and their inhibitors in tumour growth and invasion* **1999**, AnnMed, 31, 34-45.

249. Kleiner, D. E.; Stetler-Stevenson, W. G. *Matrix metalloproteinases and metastasis* **1999**, *Cancer ChemotherPharmacol*, 43, S42-S51.
250. Ray, J. M.; Stetler-Stevenson, W. G. *Gelatinase A activity directly modulates melanoma cell adhesion and spreading* **1995**, *Embo J*, 14, 908-17.
251. Hu, J.; Van den Steen, P. E.; Sang, Q. X.; Opdenakker, G. *Matrix metalloproteinase inhibitors as therapy for inflammatory and vascular diseases* **2007**, *NatRevDrug Discov*, 6, 480-98.
252. Yong, V. W. *Metalloproteinases: mediators of pathology and regeneration in the CNS* **2005**, *NatRevNeurosci*, 6, 931-44.
253. Whittaker, M.; Floyd, C. D.; Brown, P.; Gearing, A. J. *Design and therapeutic application of matrix metalloproteinase inhibitors* **1999**, *ChemRev*, 99, 2735-76.
254. Cataldo, D.; Munaut, C.; Noel, A.; Frankenne, F.; Bartsch, P.; Foidart, J. M.; Louis, R. *Matrix metalloproteinases and TIMP-1 production by peripheral blood granulocytes from COPD patients and asthmatics* **2001**, *Allergy*, 56, 145-51.
255. Mautino, G.; Oliver, N.; Chanez, P.; Bousquet, J.; Capony, F. *Increased release of matrix metalloproteinase-9 in bronchoalveolar lavage fluid and by alveolar macrophages of asthmatics* **1997**, *AmJRespirCell MolBiol*, 17, 583-91.
256. Prikk, K.; Maisi, P.; Pirila, E.; Reintam, M. A.; Salo, T.; Sorsa, T.; Sepper, R. *Airway obstruction correlates with collagenase-2 (MMP-8) expression and activation in bronchial asthma* **2002**, *LabInvest*, 82, 1535-45.
257. Coussens, L. M.; Werb, Z. *Matrix metalloproteinases and the development of cancer* **1996**, *Chemistry & Biology*, 3, 895-904.
258. Brown, P. D. *Clinical studies with matrix metalloproteinase inhibitors* **1999**, *Apmis*, 107, 174-80.
259. Denis, L. J.; Verweij, J. *Matrix metalloproteinase inhibitors: present achievements and future prospects* **1997**, *InvestNew Drugs*, 15, 175-85.
260. Wojtowicz-Praga, S. M.; Dickson, R. B.; Hawkins, M. J. *Matrix metalloproteinase inhibitors* **1997**, *InvestNew Drugs*, 15, 61-75.

261. Hidalgo, M.; Eckhardt, S. G. *Development of Matrix Metalloproteinase Inhibitors in Cancer Therapy* **2001**, J Natl Cancer Inst, 93, 178-93.
262. Mook, O. R. F.; Frederiks, W. M.; Van Noorden, C. J. F. *The role of gelatinases in colorectal cancer progression and metastasis* **2004**, Biochimica et Biophysica Acta (BBA) - Reviews on Cancer, 1705, 69-89.
263. Pirard, B. *Insight into the structural determinants for selective inhibition of matrix metalloproteinases* **2007**, Drug DiscovToday, 12, 640-6.
264. Skiles, J. W.; Gonnella, N. C.; Jeng, A. Y. *The design, structure, and clinical update of small molecular weight matrix metalloproteinase inhibitors* **2004**, CurrMedChem, 11, 2911-77.
265. Puerta, D. T.; Lewis, J. A.; Cohen, S. M. *New beginnings for matrix metalloproteinase inhibitors: identification of high-affinity zinc-binding groups* **2004**, Journal of the American Chemical Society, 126, 8388-9.
266. Puerta, D. T.; Cohen, S. M. *Examination of novel zinc-binding groups for use in matrix metalloproteinase inhibitors* **2003**, InorgChem, 42, 3423-30.
267. Puerta, D. T.; Griffin, M. O.; Lewis, J. A.; Romero-Perez, D.; Garcia, R.; Villarreal, F. J.; Cohen, S. M. *Heterocyclic zinc-binding groups for use in next-generation matrix metalloproteinase inhibitors: potency, toxicity, and reactivity* **2006**, JBiollnorgChem, 11, 131-8.
268. Farkas, E.; Katz, Y.; Bhusare, S.; Reich, R.; Roschenthaler, G. V.; Konigsmann, M.; Breuer, E. *Carbamoylphosphonate-based matrix metalloproteinase inhibitor metal complexes: solution studies and stability constants. Towards a zinc-selective binding group* **2004**, JBiollnorgChem, 9, 307-15.
269. Reich, R.; Katz, Y.; Hadar, R.; Breuer, E. *Carbamoylphosphonate matrix metalloproteinase inhibitors 3: in vivo evaluation of cyclopentylcarbamoylphosphonic acid in experimental metastasis and angiogenesis* **2005**, ClinCancer Res, 11, 3925-9.
270. Campestre, C.; Agamennone, M.; Tortorella, P.; Preziuso, S.; Biasone, A.; Gavuzzo, E.; Pochetti, G.; Mazza, F.; Hiller, O.; Tschesche, H.; Consalvi, V.; Gallina, C. *N-Hydroxyurea as zinc binding group in matrix metalloproteinase inhibition: mode of binding in a complex with MMP-8* **2006**, Bioorganic & medicinal chemistry letters, 16, 20-4.

271. Michaelides, M. R.; Dellaria, J. F.; Gong, J.; Holms, J. H.; Bouska, J. J.; Stacey, J.; Wada, C. K.; Heyman, H. R.; Curtin, M. L.; Guo, Y.; Goodfellow, C. L.; Elmore, I. B.; Albert, D. H.; Magoc, T. J.; Marcotte, P. A.; Morgan, D. W.; Davidsen, S. K. *Biaryl ether retrohydroxamates as potent, long-lived, orally bioavailable MMP inhibitors* **2001**, *Bioorganic & medicinal chemistry letters*, 11, 1553-6.
272. Auge, F.; Hornebeck, W.; Decarme, M.; Laronze, J. Y. *Improved gelatinase a selectivity by novel zinc binding groups containing galardin derivatives* **2003**, *Bioorganic & medicinal chemistry letters*, 13, 1783-6.
273. Cainelli, G.; Galletti, P.; Garbisa, S.; Giacomini, D.; Sartor, L.; Quintavalla, A. *4-Alkylidene-Azetidin-2-Ones: Novel Inhibitors of Leukocyte Elastase and Gelatinase* **2003**, *Bioorganic & medicinal chemistry*, 11, 5391-9.
274. Onaran, M. B.; Comeau, A. B.; Seto, C. T. *Squaric acid-based peptidic inhibitors of matrix metalloproteinase-1* **2005**, *JOrgChem*, 70, 10792-802.
275. Jacobsen, F. E.; Lewis, J. A.; Cohen, S. M. *A new role for old ligands: discerning chelators for zinc metalloproteinases* **2006**, *Journal of the American Chemical Society*, 128, 3156-7.
276. Lim, N. C.; Schuster, J. V.; Porto, M. C.; Tanudra, M. A.; Yao, L.; Freake, H. C.; Bruckner, C. *Coumarin-based chemosensors for zinc(II): toward the determination of the design algorithm for CHEF-type and ratiometric probes* **2005**, *InorgChem*, 44, 2018-30.
277. Puerta, D. T.; Mongan, J.; Tran, B. L.; McCammon, J. A.; Cohen, S. M. *Potent, selective pyrone-based inhibitors of stromelysin-1* **2005**, *Journal of the American Chemical Society*, 127, 14148-9.
278. Lukacova, V.; Zhang, Y.; Mackov, M.; Baricic, P.; Raha, S.; Calvo, J. A.; Balaz, S. *Similarity of binding sites of human matrix metalloproteinases* **2004**, *JBiolChem*, 279, 14194-200.
279. Cuniasse, P.; Devel, L.; Makaritis, A.; Beau, F.; Georgiadis, D.; Matziari, M.; Yiotakis, A.; Dive, V. *Future challenges facing the development of specific active-site-directed synthetic inhibitors of MMPs* **2005**, *Biochimie*, 87, 393-402.
280. Yuan, P.; Marshall, V. P.; Petzold, G. L.; Poorman, R. A.; Stockman, B. J. *Dynamics of stromelysin/inhibitor interactions studied by ¹⁵N NMR relaxation measurements: comparison of ligand binding to the S1-S3 and S'1-S'3 subsites*

- 1999, *J Biomol NMR*, 15, 55-64.
281. de Oliveira, C. A.; Zissen, M.; Mongon, J.; McCammon, J. A. *Molecular dynamics simulations of metalloproteinases types 2 and 3 reveal differences in the dynamic behavior of the S1' binding pocket* **2007**, *Curr Pharm Des*, 13, 3471-5.
 282. Lin, J. H.; Perryman, A. L.; Schames, J. R.; McCammon, J. A. *The relaxed complex method: Accommodating receptor flexibility for drug design with an improved scoring scheme* **2003**, *Biopolymers*, 68, 47-62.
 283. Adcock, S. A.; McCammon, J. A. *Molecular dynamics: survey of methods for simulating the activity of proteins* **2006**, *Chem Rev*, 106, 1589-615.
 284. Dhanaraj, V.; Williams, M. G.; Ye, Q.-Z.; Molina, F.; Johnson, L. L.; Ortwine, D. F.; Pavlovsky, A.; Rubin, J. R.; Skeeane, R. W.; White, A. D.; Humblet, C.; Hupe, D. J.; Blundell, T. L. *X-ray Structure of Gelatinase A Catalytic Domain Complexed with a Hydroxamate Inhibitor* **1999**, *Croatica chemica acta*, 72, 575-91.
 285. Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. *Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-Alkanes* **1977**, *J of Comp Physics*, 23, 327-41.
 286. Puerta, D. T.; Cohen, S. M. *Elucidating drug-metalloprotein interactions with tris(pyrazolyl)borate model complexes* **2002**, *InorgChem*, 41, 5075-82.
 287. Puerta, D. T.; Schames, J. R.; Henchman, R. H.; McCammon, J. A.; Cohen, S. M. *From Model Complexes to Metalloprotein Inhibition: A Synergistic Approach to Structure-Based Drug Discovery* **2003**, *Angewandte Chemie International Edition*, 115, 3902-4.
 288. Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I. *Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. Method* **2000**, *Journal of Computational Chemistry*, 21, 132-46.
 289. Hockney, R. W. *Potential calculation* **1968**, *Bull Am Phys Soc*, 13, 1747.
 290. Berendsen, H. J. C.; Postma, J. P. M.; Vangunsteren, W. F.; Dinola, A.; Haak, J. R. *Molecular dynamics with coupling to an external bath* **1984**, *JChemPhys*, 81, 3684.
 291. Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *A smooth particle mesh Ewald method* **1995**, *JChemPhys*, 103, 8577-93.